Eléments de Synthèse de la Parole pour PACT
Télécom ParisTech
Extraits du poly de cours de l'UE SI340

Gaël RICHARD

30 novembre 2012

Table des matières

1. Chapitre reprenant de larges extraits du polycopié de cours de T. Dutoit [11] et du cours de F. Beaugendre [3]

Chapitre I

Production et Perception de la parole

I.1 Production de la parole

La parole est le résultat acoustique résultant d'une série de mouvements des appareils respiratoires et articulatoires. De façon simple, on peut résumer le processus de production de la parole à un système dans lequel une ou plusieurs sources excitent un ensemble de cavités. La source sera soit générée au niveau des cordes vocales soit au niveau d'une constriction du conduit vocal. Dans le premier cas, la source résulte d'une vibration quasi-périodique des cordes vocales et produit ainsi une onde de débit quasi-périodique. Dans le second cas, la source sonore est soit un bruit de friction soit un bruit d'explosion qui peut apparaître s'il y a un fort rétrécissement dans le conduit vocal où si un brusque relâchement d'une occlusion du conduit vocal s'est produit. L'ensemble de cavités situées après la glotte (les cavités supraglottiques) vont ainsi être excitées par la ou les sources et "filtrer" le son produit au niveau de ces sources.

Ainsi, en changeant la forme de ces cavités, l'homme peut produire des sons différents. Les acteurs de cette mobilité du conduit vocal sont communément appelés les articulateurs.

On pourra résumer ainsi le processus de production de la parole en trois étapes essentielles :

- La génération d'un flux d'air qui va être utilisé pour faire naître une source sonore (au niveau des cordes vocales ou au niveau d'une constriction du conduit vocal : c'est le rôle de *la soufflerie*.
- La génération d'une source sonore sous la forme d'une onde quasi-périodique résultant de la vibration des cordes vocales ou/et sous la forme d'un bruit résultant d'une constriction (ou d'un brusque relâchement d'une occlusion) du conduit vocal : c'est le rôle de la *source vocale*.
- la mise en place des cavités supraglottiques (conduits nasal et vocal) pour obtenir le son désiré : c'est principalement le rôle des *différents articulateurs du conduit vocal*.

Nous détaillons dans la suite ces trois étapes du processus de production.

I.1.1 L'appareil respiratoire

L'énergie essentielle à la phonation sera produit à l'aide d'un flux d'air qui sera produit par l'appareil respiratoire (voir figure I.1). La respiration est un phénomène mécanique intégrant une phase active (l'inspiration) et une phase passive (l'expiration).

L'inspiration consiste à faire entrer de l'air dans les poumons. Pour cela, les muscles respiratoires (sterno-cleïdo-mastoïdien, scalènes, intercostaux, et surtout le diaphragme) se contractent, augmentant ainsi le volume de la cage thoracique, ce qui crée une dépression entre le feuillet pariétal de la plèvre (accroché à la cage thoracique) et le feuillet viscéral de la plèvre (accroché

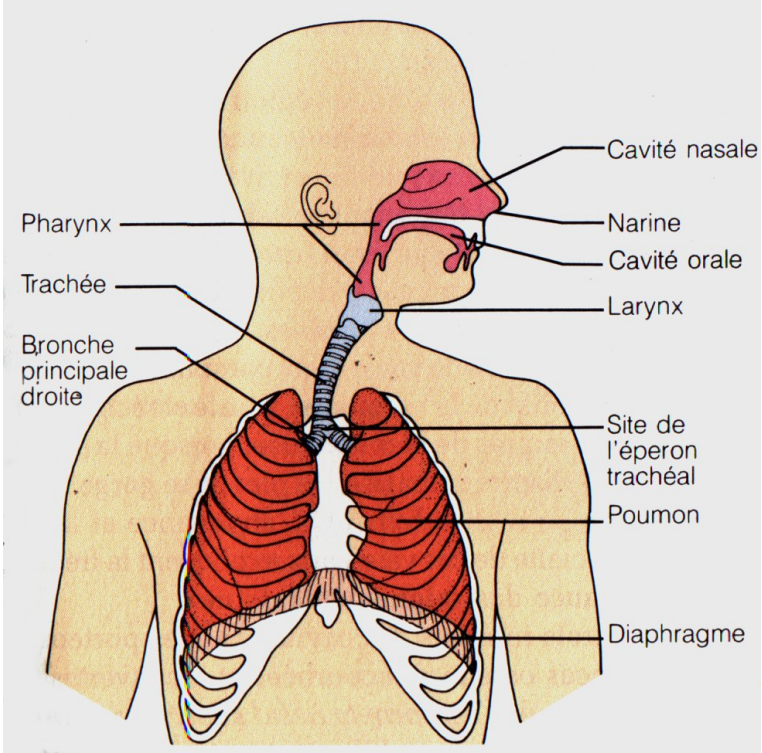


FIGURE I.1 – Schéma de l'appareil respiratoire (d'après [10])

aux poumons). Cette dépression entre les deux feuillets permet de maintenir les poumons "collés" contre les parois de la cage thoracique. L'augmentation du volume de la cage thoracique a donc augmenté le volume des poumons, ce qui a fait baisser la pression à l'intérieur des alvéoles. La pression de l'air est alors plus petite dans les poumons qu'au niveau de la bouche (qui est ouverte, donc en contact avec l'air atmosphérique) : de l'air va donc pénétrer dans les poumons pour combler la différence de pression. Il y a eu inspiration.

Contrairement à l'inspiration qui est active (c'est à dire qui met en jeu un effort musculaire) l'expiration est passive, le simple relâchement des muscles de l'inspiration permet à la cage thoracique de retrouver son volume normal (avant l'inspiration) les poumons vont donc se comprimer, entraînant une augmentation de la pression à l'intérieur des alvéoles, l'air est donc chassé vers la bouche et il y a expiration. Le cycle respiratoire peut recommencer. La fréquence respiratoire (nombre de mouvements respiratoires) est de 14 à 16 par minutes chez l'adulte (24-30/min chez l'enfant et 40-50/min chez le nouveau né).

Cependant, pour produire de la parole, et notamment pour produire de la parole forte, il est nécessaire de faire un effort musculaire supplémentaire lors de l'expiration. L'expiration de l'air n'est plus ici passive. On parlera de soufflerie.

Dans le cas d'une expiration active, c'est le diaphragme (comme pour l'inspiration) qui jouera un rôle prépondérant. Si pour la parole, cet effort se fait naturellement, il est souvent nécessaire d'apprendre à bien contrôler cette expiration à l'aide du diaphragme lorsqu'on souhaite expirer l'air avec une plus grande puissance tout en conservant une grande régularité comme cela est nécessaire pour les chanteurs ou les musiciens jouant des instruments à vent (notamment trompette, hautbois,...).

Pour plus d'information sur la respiration, on pourra consulter la description de [4].

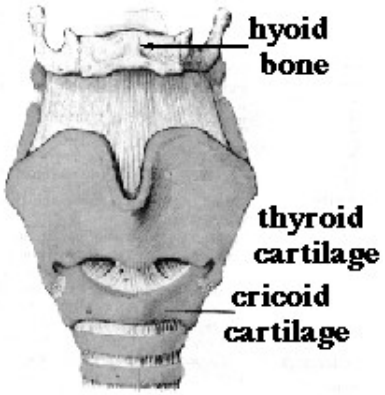


FIGURE I.2 – Schéma du larynx (d'après [23])

I.1.2 Les sources vocales

La parole est essentiellement produite par deux types de sources vocales. La première, plus sonore, est celle qui prend naissance au niveau du larynx suite à la vibration des cordes vocales. La seconde, moins sonore, prend naissance au niveau d'une constriction du conduit vocal ou lors d'un relâchement brusque d'une occlusion du conduit vocal. On parlera dans ce cas de sources de bruit.

Le larynx

Le larynx est un organe situé dans le cou qui joue un rôle crucial dans la respiration et dans la production de parole. Le larynx est plus spécifiquement situé au niveau de la séparation entre la trachée artère et le tube digestif, juste sous la racine de la langue. Sa position varie avec le sexe et l'âge : il s'abaisse progressivement jusqu'à la puberté et il est sensiblement plus élevé chez la femme. Le larynx assure ainsi trois fonctions essentielles :

- Le contrôle du flux d'air lors de la respiration
- La protection des voies respiratoires
- La production d'une source sonore pour la parole

Le larynx : un ensemble de cartilages : le larynx est constitué d'un ensemble de cartilages entourés de tissus mous (voir figure I.2). La partie la plus proéminente du larynx est formée du thyroïde. La partie antérieure de cartilage est communément appelée la "pomme d'Adam". On trouve juste au dessus du larynx un os en forme de 'U' appelé l'os hyoid. Cette os relie le larynx à la mandibule par l'intermédiaire de muscles et de tendons qui joueront un rôle important pour élever le larynx pour la déglutition ou la production de parole.

La partie inférieure du larynx est constituée d'un ensemble de pièces circulaires : le cricoïde sous lequel on trouve les anneaux de la trachée artère.

Au centre du larynx, on trouve les cordes vocales (on parlera aussi couramment de la glotte pour désigner l'ensemble constitué des cordes vocales, même si rigoureusement la glotte désigne plutôt l'espace se trouvant entre les cordes vocales). Les cordes vocales sont particulièrement

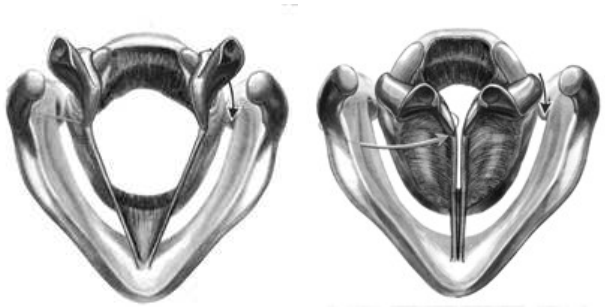


FIGURE I.3 – Les cordes vocales en position ouvertes durant la respiration (à gauche) et fermées pour la production de parole (à droite),(d'après [23])

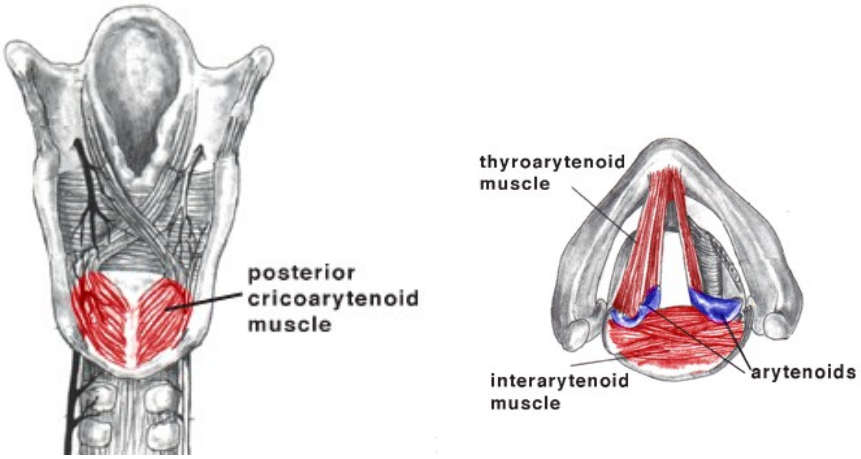


FIGURE I.4 – Schéma des muscles intrinsèques du larynx (d'après [23])

importantes puisqu'elles jouent un rôle fondamental dans les trois fonctions essentielles du larynx.

Les cordes vocales sont constituées de muscles recouverts d'un tissus assez fin couramment appelé la muqueuse. Sur la partie arrière de chaque corde vocale, on trouve une petite structure faite de cartilages : les aryténoides. De nombreux muscles y sont rattachés qui permettent de les écarter pour permettre la respiration. Durant la production de parole, les aryténoides sont rapprochés (voir figure I.3). Sous la pression de l'air provenant des poumons, les cordes vocales s'ouvrent puis se referment rapidement. Ainsi, lorsqu'une pression soutenue de l'air d'expiration est maintenue, les cordes vocales vibrent et produise un son qui sera par la suite modifié dans le conduit vocal pour donner lieu à un son voisé. Ce processus de vibration des cordes vocales est décrit un peu plus en détail ci-dessous.

Les muscles du larynx Les mouvements du larynx sont contrôlés par deux groupes de muscles. On distingue ainsi les muscles intrinsèques (ceux qui contrôlent le mouvement des cordes vocales et des muscles à l'intérieur du larynx) et les muscles extrinsèques (qui contrôlent la position du larynx dans le cou).

La figure I.4 montre les muscles intrinsèques. Les cordes vocales sont ouvertes par une paires de muscles (le muscle cricoaryténoide postérieur) qui sont situés entre la partie arrière du cricoïde et le cricoarytenoïd.

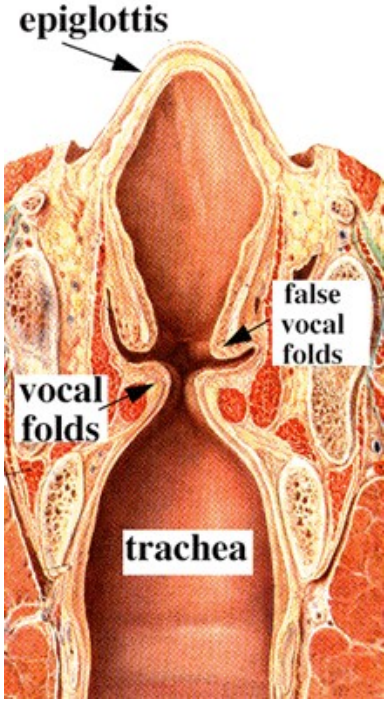


FIGURE I.5 – Vue longitudinale du larynx (d'après [23])

Plusieurs muscles aident pour fermer et tendre les cordes vocales. Les cordes vocales sont elles-même constituées d'un muscle, le thyroaryténoide. Un autre muscle, l'interaryténoide , permet de rapprocher ces deux cartilages. Le muscle cricoaryténoide latéral qui est lui aussi situé entre l'aryténoide et le cartilage cricoïde sert à la fermeture du larynx.

Le muscle cricothyroide va du cartilage cricoide jusqu'au cartilage thyroide. Lorsqu'il se contracte, le cartilage cricoid bascule en avant et tend les cordes vocales ce qui résultera à un élèvement de la voix.

Les muscles extrinsèques n'affectent pas le mouvement des cordes vocales mais élèvent ou abaissent le larynx dans sa globalité.

Description détaillée de la phonation La figure I.5 donne une vue schématique d'une coupe verticale du larynx. Sur ce schéma, les cordes vocales sont ici clairement séparées, comme elles seraient durant la respiration. On peut également remarquer au-dessus des cordes vocales, des tissus ayant pour principal rôle d'éviter le passage de substances dans la trachée durant la déglutition : ce sont les fausses cordes vocales. Il est important de noter qu'elles ne jouent aucun rôle lors de la phonation. Le cartilage mou en forme grossière de langue qui se trouve au-dessus est appelé l'épiglotte et a également un rôle pour protéger l'accès de la trachée lors de la déglutition.

Lors de la phonation, les cordes vocales sont tout d'abord rapprochées l'une de l'autre par les muscles du larynx. Lorsqu'elles sont fermées, l'action des muscles respiratoires font augmenter la pression subglottique (juste en dessous des cordes vocales). Lorsque cette pression est supérieure à celle forçant les cordes vocales l'une contre l'autre, une bouffée d'air s'échappe à travers les cordes vocales qui se sont alors momentanément ouvertes. Ensuite, deux forces vont concourir à les rapprocher : leur élasticité et l'effet d'aspiration provoqué par le passage de l'air au niveau

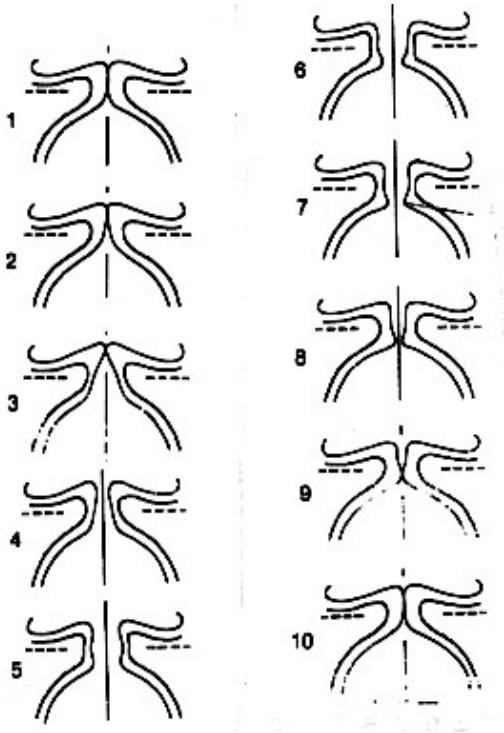


FIGURE I.6 – Schéma de vibration des cordes vocales (d'après [23])

de la glotte (en raison de l'effet Bernouilli). La pressure subglottique augmente de nouveau et le processus se répète. On parlera ainsi dans ce cas de vibration des cordes vocales. Il est important de remarquer que les cordes vocales ne produisent pas un son en vibrant comme le ferait une corde de guitare, mais qu'elles produisent un son en créant des bouffées d'air qui impliquent un changement de pression d'air de façon quasi périodique.

Le diagramme ci-dessus (figure I.6) montre une vue schématique d'une section longitudinale de l'ouverture et fermeture des cordes vocales. On peut voir que ces dernières ne s'ouvrent pas uniformément, mais vont d'abord se séparer par leur base. De même, sous l'effet Bernoulli, les cordes vocales se refermeront d'abord par leur base et seulement ensuite sur toute leur hauteur.

On pourra trouver des détails supplémentaires dans [13] ou [7].

Les sources de bruit

Les sources de bruits peuvent apparaître soit dans le larynx, soit dans le conduit vocal, soit encore dans les deux à la fois. Nous allons décrire ci-dessous les principaux moyens de générer du bruit (ou plus précisément un signal aléatoire) à l'aide de notre appareil phonatoire, en nous restreignant aux bruits existants dans la langue française.

On peut distinguer différents types de bruits suivant leurs modes de phonation :

- la première situation est rencontrée lorsque les cordes vocales sont écartées et ne vibrent pas. Le bruit ne pourra donc naître que dans le conduit vocal.
  Ces bruits sont produits suite à une obstruction suffisamment étroite du conduit vocal, réalisée par exemple en rapprochant la langue du palais ou des dents :
  - *les bruits fricatifs* où l'obstruction du conduit vocal n'est que partielle ce qui a pour conséquence de générer un bruit turbulent au point de constriction.

- *les bruits d'explosion* qui naissent suite à l'ouverture brutale d'une obstruction totale du conduit vocal. Le bruit est alors constitué de deux composantes : 1) un bruit impulsif causé par le relâchement soudain de la pression d'air suivi 2) d'un bruit d'aspiration [1] causé par turbulence à travers la constriction prés du point d'articulation (bruit similaire au bruit fricatif mais de durée moindre).
- tous les *bruits de bouches* tels les claquements de langue, ou bruits de lèvres mais qui ne jouent pas de rôle linguistique.

– la seconde situation est celle rencontrée pour la voix chuchotée pour laquelle la source de bruit se situe au niveau de la glotte. Ici, les cordes vocales sont rapprochées, mais les aryténoïdes sont écartés et un bruit de friction va donc naître dans cette ouverture. On peut également ranger dans cette catégorie, les bruits produits par occlusion glottale. Dans ce cas, on aura un relâchement d'air comme pour les plosives, mais l'obstruction étant ici au niveau de la glotte.

I.1.3 Les cavités supraglottiques

Il existe 2 cavités supraglottiques (v. figure I.7) : *le conduit nasal* (ou fosses nasales) et *le conduit vocal*.

Le conduit vocal peut être vu comme un tube acoustique de section variable. Il s'étend de la glotte (l'espace situé entre les cordes vocales) jusqu'aux lèvres. Pour un adulte, le conduit vocal mesure environ 17 cm. La forme du conduit vocal varie en fonction du mouvement des articulateurs qui sont les lèvres, la mâchoire, la langue et le velum. Ces articulateurs sont brièvement décrits ci-dessous.

Le conduit nasal est un passage auxiliaire pour la transmission du son. Il commence au niveau du velum et se termine aux fosses nasales. Pour un homme adulte, cette cavité mesure environ 12 cm et possède un volume d'environ 60 cm^3. Le couplage acoustique entre les deux cavités est contrôlé par l'ouverture au niveau du velum (Sur la figure I.7, on notera que le velum -ou voile du palais- est largement ouvert. Dans ce cas, on aura la production d'un son nasal. Dans le cas contraire, lorsque le velum ferme le conduit nasal le son produit sera dit non-nasal.

Sachant que l'on ne peut pas vraiment contrôler la forme du conduit nasal, nous restreindrons la description plus détaillé aux articulateurs du conduit vocal.

La langue

La langue est une structure frontière, appartenant à la fois à la cavité buccale pour sa partie dite mobile et au glosso-pharynx pour sa partie dite fixe [4].

La langue mobile a la forme d'une pyramide à faces arrondies, constituée d'une charpente musculaire, pouvant se rétracter ou s'étendre dans toutes les dimensions jusqu'à sa pointe et se tourner dans toutes les directions. Elle est revêtue sur sa face dorsale d'un tapis de papilles (les papilles gustatives). Ses bords latéraux effleurent les dents latérales tandis que sa pointe vient affleurer les dents antérieures de la mandibule. Elle trouve sa limite postérieure au niveau d'une rangée de grosses papilles, sans rôle particulier, disposées en V à pointe postérieure, le V lingual, qui la sépare arbitrairement de la base de langue. Il n'y a pas de différence de structure notable sur le plan musculaire entre ces 2 parties, que l'on distingue pour une question anatomique par leur condition de mobilité. Outre sa fonction gustative, cette partie mobile joue un rôle essentiel dans la mastication, la déglutition et, bien sur, l'articulation des sons. La langue appliquée contre

1. Notons que le terme bruit d'aspiration est parfois réservé au bruit émis au niveau (ou prés) de la glotte ([26], [12]).

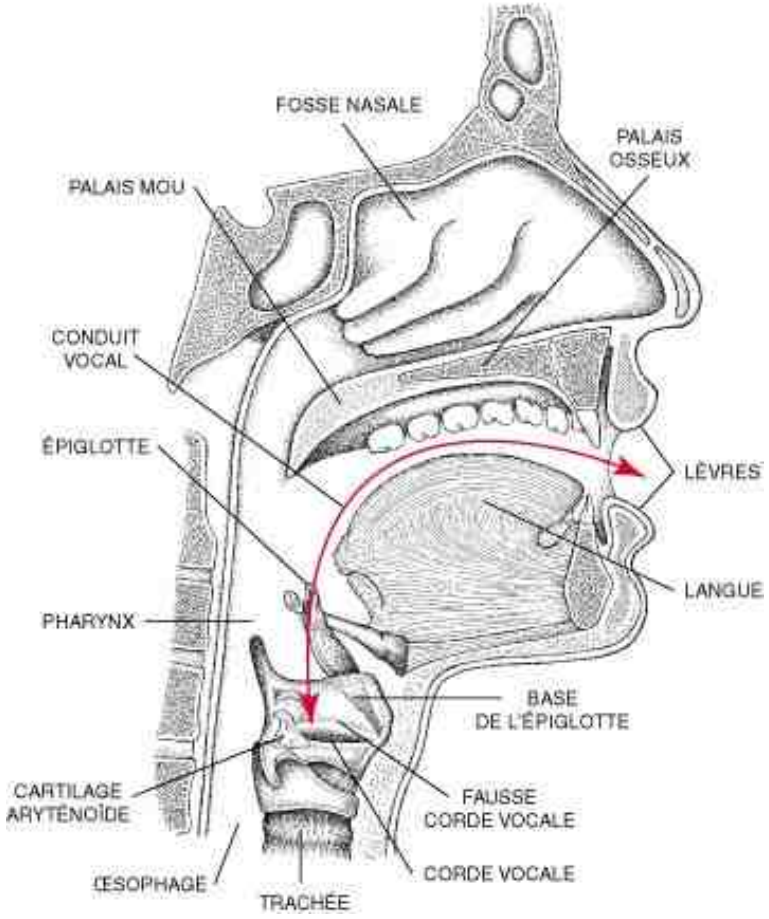


FIGURE I.7 – Vue schématique du conduit vocal humain d'après [19]

le palais ou les dents constituent un organe vibratoire accessoire, intervenant dans la formation des consonnes.

La base de la langue, formant la pente pharyngée, est la partie peu mobile postérieure de la langue et se raccorde dans sa partie basse à l'épiglotte. Sa masse musculaire large est assise sur l'os hyoïde sur lequel elle s'insère en partie en arrière, ses attaches antérieures se faisant sur la face interne des angles mandibulaires. Elle a de également l'importance pour la phonation. (Pour obtenir plus de précision sur la langue on pourra consulter [4] dont la description ci-dessus est extraite).

On comprend que la langue est un articulateur fondamental puisque sa position est déterminante dans le conduit vocal.

La mâchoire

La mâchoire possède un nombre de degrés de liberté plus faible et étant un corps rigide ne peut pas se déformer comme la langue. Néanmoins, la mâchoire peut non seulement s'ouvrir et se fermer, mais peut également s'avancer ou effectuer des mouvements de rotation (d'amplitude toutefois assez modérée). Son rôle dans la parole n'est cependant pas primordial dans la mesure ou il est possible en bloquant la mâchoire de parler de façon très intelligible. On verra toutefois que la modélisation articulatoire de la mâchoire présente un intérêt pour la synthèse de visage parlants naturels.

Les lèvres

Les lèvres sont situées à l'extrémité du conduit vocal et comme pour la langue, elles possèdent une grande mobilité en raison des nombreux muscles impliqués dans leur contrôle. Les points de jonction des lèvres supérieure et inférieure s'appellent les commissures et jouent un grand rôle dans la diplomatie (pour le sourire, bien sur...). Au point de vue acoustique, c'est l'espace intérolabial qui est important. On peut observer différents mouvements importants pour la phonation dont :

- l'occlusion (les lèvres sont fermées)
- la protrusion (les lèvres sont avancées vers l'avant)
- l'élévation et l'abaissement de la lèvre inférieure
- l'étirement, l'abaissement ou l'élévation des commissures

I.2 Les sons de la parole vus sous une approche production

Nous allons voir dans cette partie comment on peut classer les sons suivant leur mode de production. La parole, qu'elle qu'en soit la langue, est constituée d'un nombre finis d'éléments sonores distinctifs. Ces éléments forment les unités linguistiques élémentaires et ont la propriété de changer le sens d'un mot. Ces unités élémentaires sont appelés *phonèmes*. Une définition du phonème peut ainsi être énoncée sous la forme : "Les phonèmes sont les éléments sonores les plus brefs qui permettent de distinguer différents mots"

Les phonèmes peuvent ainsi être vus comme les éléments de base pour le codage de l'information linguistique. L'étude des sons du langage est souvent divisée en deux approches :

- *La phonétique* qui s'intéresse à la manière dont les sons de parole sont produits, transmis et perçus.
- *La phonologie* qui s'intéresse à découvrir comment ces sons participent au fonctionnement de la langue dans l'acte de parole et à son codage.

Il est parfois difficile de comprendre la subtile différence entre ces deux approches. L'exemple du /r/ en français est souvent donné car il permet de mieux saisir cette différence. Lorsque le mot "rocailleux" est prononcé, il peut l'être soit avec un [r] roulé (produit avec le bout de la langue) soit avec un [r] grasseyé (produit avec le dos de la langue dans la gorge). Ces deux prononciations ne provoquent pas de changement de sens, mais les deux [r] sont pourtant bien différents du point de vue de la production. On dira qu'ils sont phonétiquement distincts et phonologiquement semblables.

Dans ce document, nous ne donnerons pas de description très détaillée de la phonétique ou de la phonologie. On pourra pour cela se reporter à [7] et aux nombreuses références s'y trouvant (p14). Nous allons par contre, nous attacher à décrire les différentes classes de sons en expliquant, du point de vue de la production comment ces sons sont produits. Nous commencerons cela par une brève présentation des sons du français et de la phonétique.

Notions de phonétique

La phonétique est l'un des domaines importants du traitement de la parole. Comme il est déjà indiqué ci-dessus, la phonétique s'intéresse à comprendre la façon dont les sons sont produits et perçus. Nous avons déjà parlé des phonèmes qui sont les éléments sonores les plus brefs d'une langue.

Cependant, ces phonèmes peuvent se regrouper en classes dont les éléments partagent des caractéristiques communes. On parlera ici de "traits distinctifs". Un trait distinctif sera ainsi l'expression d'une similarité au niveau articulatoire, acoustique ou perceptif des sons concernés.

Par exemple, pour les voyelles on distinguera 4 traits distinctifs :

- *La nasalité :* la voyelle a été prononcée à l'aide du conduit vocal et du conduit nasal suite à l'ouverture du velum
- *Le degré d'ouverture* du conduit vocal
- *La position de la constriction principale* du conduit vocal, cette constriction étant réalisée entre la langue et le palais.
- *la protrusion des lèvres.*

De même, les consonnes seront classées à l'aide de 3 traits distinctifs :

- Le voisement : la consonne a été prononcée avec une vibration des cordes vocales
- le mode d'articulation (on distinguera les modes occlusif, fricatif, nasal, glissant ou liquide).
- La position de la constriction principale du conduit, souvent appelée lieu d'articulation qui contrairement aux voyelles n'est pas nécessairement réalisé avec le corps de la langue.

Il existe d'autres façons d'organiser les sons par exemple en opposant les sons sonnants (voyelles, consonnes nasales, liquides ou glissantes) aux sons obstruants (occlusives, fricatives).

En fait, les phonèmes (qui peuvent être décrits suivant leurs traits distinctifs) sont des éléments abstraits associés à des sons élémentaires. Bien entendu, les phonèmes ne sont pas identiques pour chaque langue et le /a/ du français (comme par exemple dans "Paris") n'est pas totalement équivalent au /a/ de l'anglais (par ex. dans 'cat'). Ainsi, est née l'idée de définir un alphabet phonétique international (alphabet IPA) qui permettrait de décrire les sons et les prononciations de ces sons de manière compacte et universelle.

On trouvera de plus amples informations sur le site de l'IPA (voir [16]) dont a été extrait le tableau complet de l'alphabet phonétique international donné figure I.8 :

On pourra noter que les symboles phonétiques utilisés pour le français sont un sous-ensemble de l'alphabet phonétique international.

Nous allons voir ci-dessous de manière un peu plus précise, les caractéristiques de chaque classe de sons.

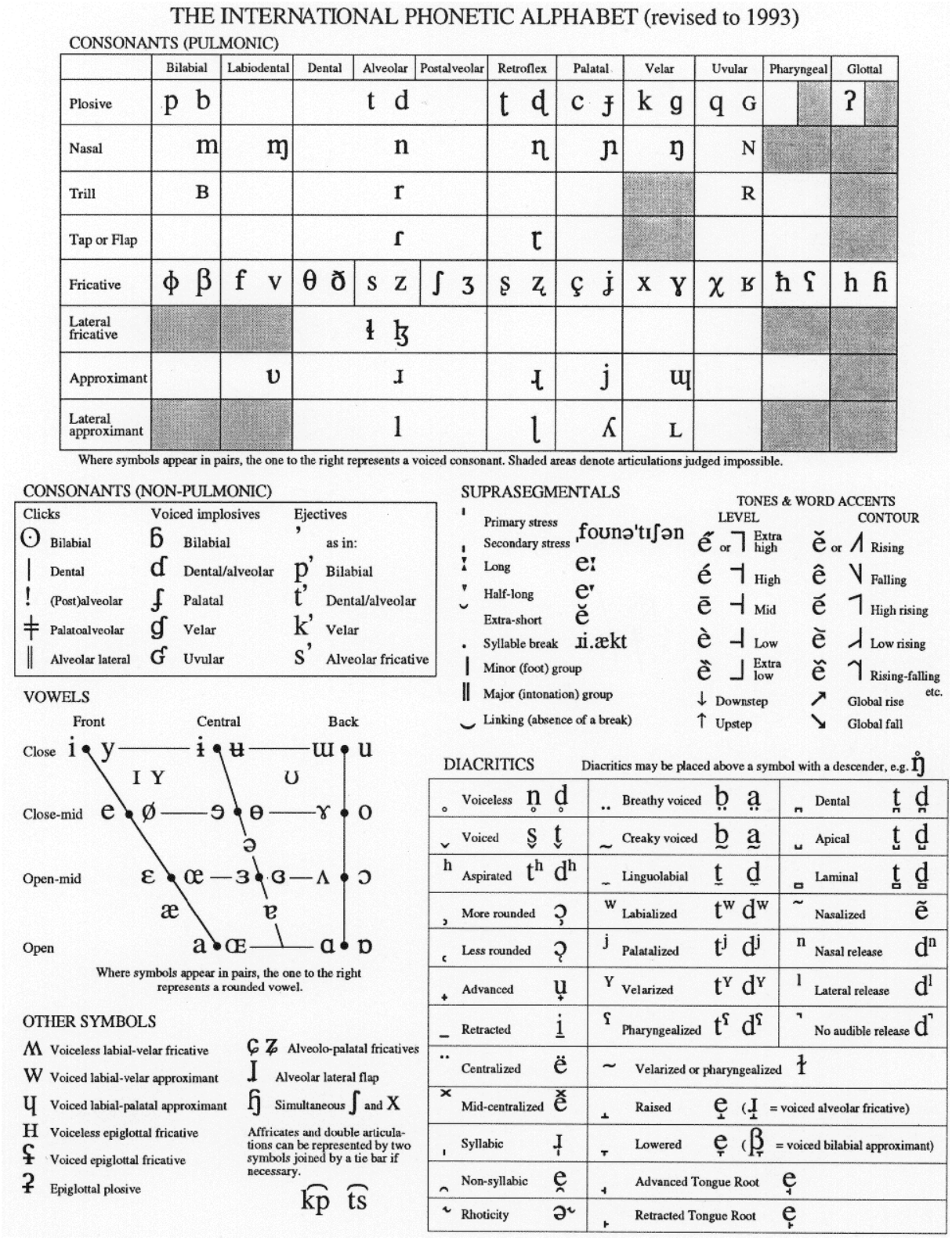


FIGURE I.8 – Tableau complet de l'alphabet phonétique international [16]

Les voyelles

Les voyelles sont typiquement produites en faisant vibrer ses cordes vocales. Le sons de telle ou telle voyelle est alors obtenu en changeant la forme du conduit vocal à l'aide des différents articulateurs. Dans un mode d'articulation normal (sans articulation exagérée), la forme du conduit vocal est maintenue relativement stable pendant quasiment toute la durée de la voyelle. Comme nous l'avons vu, ci-dessus les voyelles seront caractérisées par quatre principaux traits distinctifs.

- **les voyelles antérieures/postérieures** Ainsi, en référence au lieu de la principale constriction du conduit vocal (qui sera réalisé par la position du corps de la langue) on parlera de voyelles antérieures, centrales et postérieures. Ainsi, pour une voyelle postérieure (comme /u/ dans "houx"), le corps de la langue sera placé très en arrière du conduit vocal, alors que pour une voyelle antérieure (comme /i/ dans "lit"), le corps de la langue sera ramené vers les dents.
- **les voyelles ouvertes et fermées** en référence à l'ouverture du conduit vocal, on parlera de voyelles ouvertes ou fermées. Ainsi, pour une voyelle fermée (comme /i/ dans "lit"), on aura un conduit vocal avec une importante constriction ce qui fera souvent naître un léger bruit de chuintement supplémentaire. Cette forme du conduit vocal correspond à une position haute de la langue. Pour une voyelle ouverte, à l'inverse, on aura une position de la langue plus basse et ainsi une constriction moins importante (comme /a/ dans "patte")
- **les voyelles arrondies** en référence à la protrusion des lèvres, on parlera de voyelles arrondies (ou labialisées) lorsqu'elles sont prononcées en avançant les lèvres vers l'avant (comme pour le son /u/ dans "houx"). A l'opposée, on trouve des voyelles non-arrondies (telles que le /i/ dans "lit") qui sont prononcées en étirant les lèvres.
- **les voyelles nasales** Certaines voyelles mettent également en jeu le conduit nasal dont l'excitation est rendue possible grâce à l'abaissement du voile du palais. On les appellera les *voyelles nasales*. C'est notamment le cas de /an/ dans "pente".

Ainsi, pour caractériser une voyelle on pourra la décrire à l'aide des traits ci-dessus. Par exemple, la voyelle /i/ de "lit" est antérieure, fermée, non arrondies et non nasale. On trouvera plus d'informations dans par exemple Ladefodged51 et Malmberg79

Le tableau donné figure I.9 donne une classification des phonèmes du français suivant ces traits distinctifs généraux.

Les consonnes

Comme pour les voyelles, les consonnes vont pouvoir être regroupées en traits distinctifs. Contrairement aux voyelles par contre, elles ne sont pas exclusivement voisées (même si les voyelles prononcées en voix chuchotée sont, dans ce cas également, non voisées) et ne sont pas nécessairement réalisées avec une configuration stable du conduit vocal.

Les consonnes voisées On parlera de consonnes voisées lorsqu'elles auront été produites avec une vibration des cordes vocales (comme par exemple /b/ dans "bol" où les cordes vocales vibrent avant le relâchement de la constriction). Lorsqu'en plus du voisement, une source de bruit est présente due à une constriction du conduit vocal, on pourra parler de consonnes à excitation mixte (c'est le cas par exemple du /v/ dans "vent").

Les fricatives elles sont produites par un flux d'air turbulent prenant naissance au niveau d'une constriction du conduit vocal. On distingue plusieurs fricatives suivant le lieu de cette constriction principale :

CONSONNES Mode d'articulation ↓	Labiales	Dentales	Vélo-palatales ←	Lieu d'articulation
Occlusives				
non voisées	[p]	[t]	[k]	
voisées	[b]	[d]	[g]	
Nasales	[m]	[n]	[ɲ]	
Fricatives				
non voisées	[f]	[s]	[z]	
voisées	[v]	[z]	[ʒ]	
Glissantes	[w]	[ɥ]	[j]	
Liquides		[l]	[R]	
VOYELLES				
Orales	Antérieures		Postérieures	
	Non arrondies		Arrondies	
Fermées	[i]	[y]	[u]	
	[e]	[ø]	[o]	
	[ɛ]	[œ]	[ɔ]	
Ouvertes	[a]			
Nasales	Antérieures		Postérieures	
Fermées	[ɛ̃]		[õ]	
Ouvertes		[ã]		

FIGURE I.9 – Classification des phonèmes du français [7]

- Les labio-dentales, pour une constriction réalisée entre les dents et les lèvres (comme pour le /f/ dans "foin")
- Les dentales, pour une constriction au niveau des dents (comme pour le /θ/ anglais dans "thin")
- Les alvéolaires, pour une constriction juste derrière les dents (comme pour le /s/ dans "son")
- Les palatales, pour une constriction au niveau du palais dur (comme pour le /$\int$/ dans chat).
- Les laryngales, pour une excitation au niveau de la glotte (comme pour le /h/ anglais dans "he")

En fait, suivant les langues, en regardant plusieurs langues, on s'aperçoit que quasiment tous les points d'articulations du conduit vocal peuvent être utilisés pour réaliser des fricatives. C'est d'ailleurs l'une des difficultés de l'apprentissage des langues étrangères car il n'est pas aisé d'apprendre à réaliser des sons qui demande de positionner la langue à des endroits inhabituels (par exemple la dorso-vélaire allemande /ch/ de "ich", la palatale suédoise rencontrée dans le mot "sju", 7 en français qui est réalisée avec une constriction située entre le /s/ et le /$\int$/ français, etc...)

les plosives Elles sont caractérisées par une dynamique importante du conduit vocal. Elles sont réalisées en fermant le conduit vocal en un endroit. L'air provenant des poumons crée alors une pression derrière cette occlusion qui est ensuite soudainement relâchée suite au mouvement rapide des articulateurs ayant réalisé cette occlusion. De même, que pour les fricatives, l'un des traits distinctifs entre les plosives est le lieu d'articulation. Pour les plosives, on aura ainsi :

- Les labiales, pour une occlusion réalisée au niveau des lèvres (comme pour le /p/ dans "par")
- Les dentales, pour une occlusion au niveau des dents (comme pour le /t/ dans "tarte"). Notons qu'en anglais le /d/ ou le /t/ seront articulés un peu plus en arrière et on parlera alors de plosives alvéolaires.
- Les vélo-palatales, pour une occlusion au niveau du palais (comme pour le /k/ dans "cake").

En plus du lieu d'articulation, les plosives peuvent également être voisées ou non voisées. Ainsi, une dentale voisée (/d/) se distinguera uniquement par la présence de voisement (vibration des cordes vocales) du /t/ qui est prononcée avec le même lieu d'articulation.

les consonnes nasales Elles sont en général voisées et sont produites en effectuant une occlusion complète du conduit vocal et en ouvrant le vélum permettant au conduit nasal d'être l'unique résonateur. Comme pour les autres consonnes, on aura, suivant le lieu d'articulation :

- Les labiales, pour une occlusion du conduit vocal réalisée au niveau des lèvres (comme pour le /m/ dans "main")
- Les dentales, pour une occlusion du conduit vocal au niveau des dents (comme pour le /n/ dans "non"). Notons qu'en anglais le /n/ sera articulé un peu plus en arrière et on parlera alors plutôt de nasales alvéolaires.
- Les vélo-palatales, pour une occlusion du conduit vocal au niveau du palais (comme pour le /ŋ/ dans "parking").

Les glissantes et les liquides cette classe de consonnes regroupe des sons qui ressemblent aux voyelles. Les liquides sont d'ailleurs parfois appelées semi consonnes ou semi-voyelles. Les glissantes et les liquides, en général, voisées et non nasales. Les glissantes, comme leur nom l'indique, sont des sons en mouvement et précédent toujours une voyelle (ou un son vocalique). On aura :

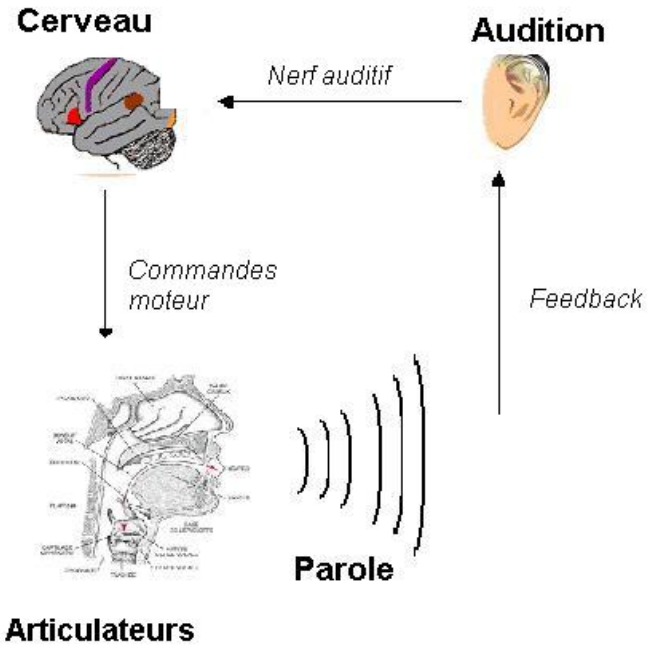


FIGURE I.10 – Système de production et feedback auditif

– la glissante vélo-palatales /R/ comme dans "rat"
– la dentale /l/ comme dans "lit".

Les liquides (ou semi-voyelles) sont des sons tenus, très similaires aux voyelles mais en général avec une constriction plus conséquente et avec l'apex de la langue plus relevé. On aura :

– la labiale "Wé", noté /w/ que l'on trouve dans "loi" pour former le son s'intercalant entre le /l/ et le /a/.
– la dentale "Ué", noté /y/, que l'on trouve dans "nuit" pour former le son s'intercalant entre le /u/ et le /i/. En français, ce son est toujours suivi du phonème /i/.
– la vélo-palatale ("yod") comme /j/ pour former le son "ill" entre le /i/ et le /e/ dans "piller".

I.3 Notions de perception des sons de parole

Les sons de la parole ont été présentés sous l'angle de la production. Cependant, la production ne peut pas être totalement dissocié de la perception. En effet, à la base, la parole est produite dans le but d'être écouté (et comprise même si certains parlent parfois pour ne rien dire..). Ainsi, la production de parole est en fait contrôlée par ce que l'on entend (voir figure I.10) et on peut ainsi voir le mécanisme de production comme un système à boucle de retour ("feedback"). Ce mécanisme de feedback est réellement important et cela est mis en évidence chez les personnes qui ont perdus l'audition. En effet, au bout d'un certain laps de temps (quelques années) leur parole se détériore significativement.

Cette brève introduction montre l'importance de la perception dans cette chaîne de la parole. Nous allons rappeler ci-dessous quelques éléments de perception des sons en précisant les aspects

importants de cette perception dans le cas d'un signal de parole.

I.3.1 Éléments de perception

La perception d'un son de parole est généralement séparée en deux phases principales :
- La transmission du message acoustique (le son) au cerveau
- L'interprétation du message linguistique lié au signal acoustique reçu

La deuxième phase de ce processus est mal connue car son étude est particulièrement complexe. Au niveau du cerveau, on sait cependant que les aires de Broca et de Wernicke sont importantes pour la perception (et la production de parole). Par exemple, des lésions de l'aire de Wernicke font perdre la capacité de comprendre la parole, mais ne font pas perdre la capacité de prononcer clairement des mots ou phrases même si ceux ci sont prononcés sans aucun lien entre eux. Ainsi, l'aire de Wernicke renferme l'information nécessaire pour arranger les mots appris et former des phrases parlées ayant un sens. L'aire de Broca renferme l'information nécessaire pour la production de parole. L'aire de Broca est responsable du mouvement des articulateurs actifs lors de la production de parole (lèvres, langues, muscles de la parole). ([9])

La première phase de ce processus est elle mieux connue. Sans rentrer dans les détails rappelons que :
- L'oreille est séparée en 3 parties principales :
  - l'oreille externe allant du pavillon au tympan et réalisant une conduction aérienne.
  - L'oreille moyenne, constituée de 3 osselets (le marteau, l'enclume et l'étrier) s'étend du tympan à la fenêtre ovale et réalise une adaptation d'impédence pour transmettre les ondes acoustiques aériennes reçues au niveau de l'oreille externe vers l'oreille interne.
  - L'oreille interne dans laquelle se trouve la cochlée. La cochlée joue un rôle primordial dans la perception des sons. En effet, un son parvenant au pavillon de l'oreille sera transformé en vibration au niveau de l'entrée de la cochlée (fenêtre ovale). En fonction de sa fréquence, la vibration à un effet maximal (résonance) en un point différent de la membrane basilaire : c'est la tonotopie passive. Il est alors clair que les fréquences d'un son représenteront une information particulièrement importante pour son identification/classification.
- La sélectivité en fréquence est plus grande dans le grave que dans l'aigu. C'est cette caractéristique qui justifiera l'utilisation d'échelle Bark, ou échelles Mel pour la paramétrisation du signal de parole.
- Une oreille humaine performante perçoit des fréquences comprises entre 20 Hz (fréquence la plus grave) et 20 000 Hz (fréquence perçue la plus aiguë).

I.3.2 Description du signal de parole

Description temporelle

Le signal de parole est un signal quasi-stationnaire, c'est à dire que ses caractéristiques statistiques changent peu sur des périodes de temps suffisamment courtes (qui varieront en moyenne entre 5 et 100 ms suivant les sons). Cependant, sur un horizon de temps supérieur, il est clair que les caractéristiques du signal évolue significativement en fonction des sons prononcés.

La première approche pour étudier le signal de parole consiste à observer la forme temporelle du signal. On peut à partir de cette forme temporelle en déduire un certain nombre de caractéristiques qui pourront être utilisées pour le traitement de la parole. Il est, par exemple, assez clair de distinguer les parties voisées (dans lesquelles on peut observer une forme d'onde quasi-périodique) des parties non voisées (dans lesquelles un signal aléatoire de faible amplitude

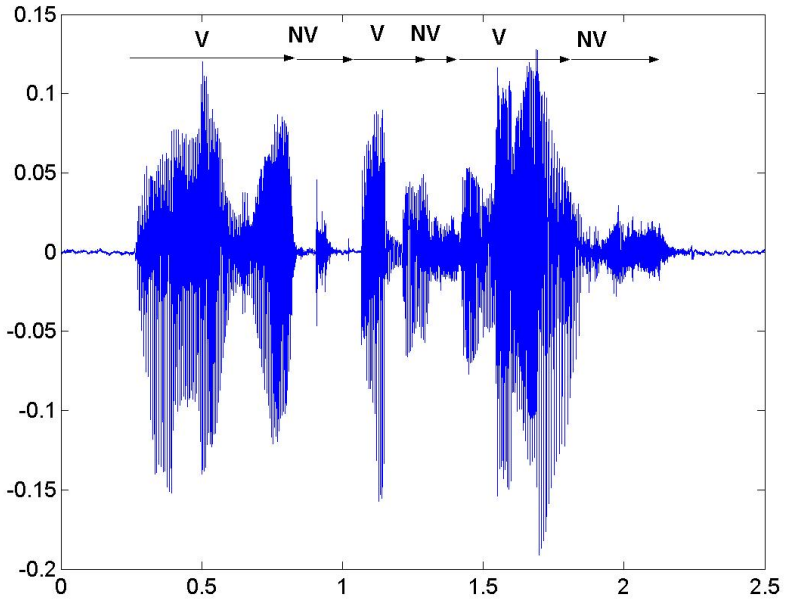


FIGURE I.11 – Signal temporel de la phrase "La musique adoucit les moeurs" : (V=partie voisée ; NV = partie non voisée

est observé). De même, on peut voir que les petites amplitudes sont beaucoup plus représentées que les grandes amplitudes ce qui pourra justifier des choix fait en codage de la parole.

Cependant, si cette segmentation apparaît assez claire sur le signal donné figure I.11, ce ne sera pas toujours le cas. Il sera, en pratique, souvent difficile de distinguer une partie non voisée prononcée faiblement du silence (surtout en présence de bruit de fond) voire de distinguer une partie voisée prononcée faiblement des parties non voisées. De plus, une telle représentation ne permet pas d'identifier/repérer les voyelles entres elles.

Description fréquentielle

Une seconde approche pour caractériser et représenter le signal de parole est d'utiliser une représentation spectrale. Clairement, la représentation la plus répandue est le *spectrogramme*. Le spectrogramme permet de donner une représentation tridimensionnelle d'un son dans laquelle l'énergie par bande de fréquences est donnée en fonction du temps.

Plus précisément, le spectrogramme représente le module de la transformée de Fourier discrète calculé sur une fenêtre temporelle plus ou moins longue. La transformée de Fourier discrète (TFD) $X(k)$ de la ième fenêtre de signal de parole $x(n)$ est donnée par [2] :

$$X_i(k) = \sum_{n=0}^{N-1} x(n) e^{-2j\pi kn/N} \tag{I.1}$$

2. notons que $x(n)$ représente en fait la version échantillonnée de x(t) aux instants nT. Pour une plus grande lisibilité, on ne conservera que l'indice n pour représenter les échantillons successifs du signal x

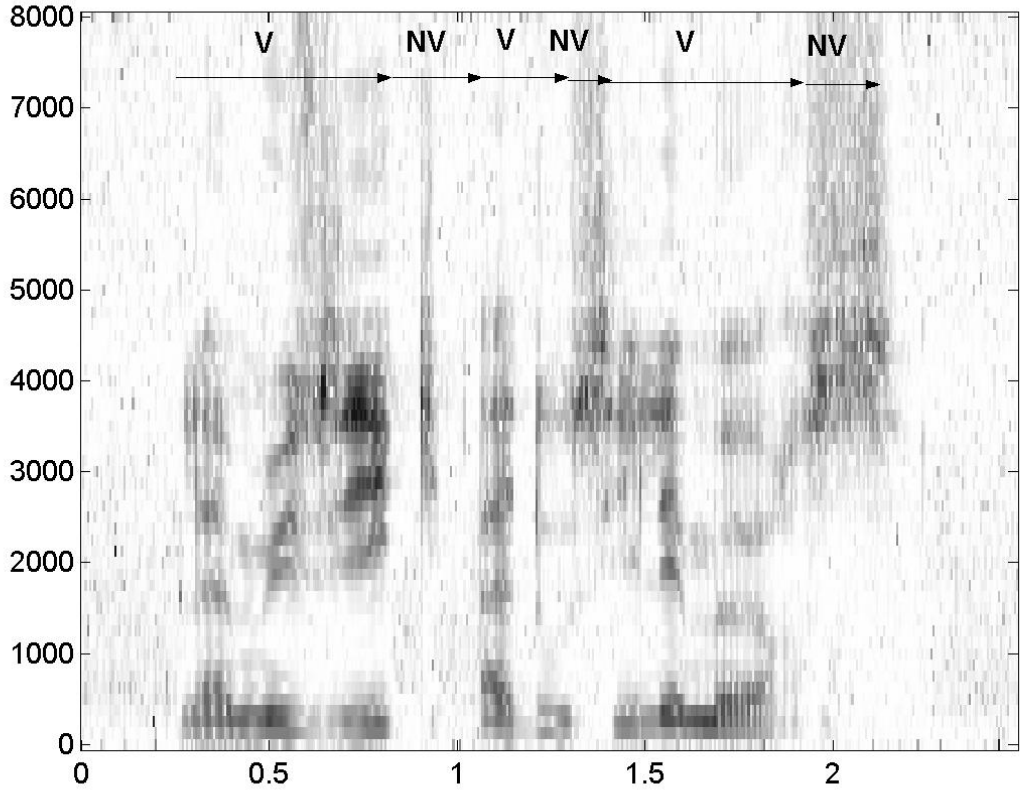


FIGURE I.12 – Spectrogramme de la phrase "La musique adoucit les moeurs" (Le spectrogramme représente le module de la transformée de Fourier où cours du temps avec les Fréquences en ordonnée, le temps en abscisse et l'énergie en niveau de gris. Ainsi une zone sombre, indique une forte énergie à la fréquence et au temps correspondants

Le spectrogramme est ensuite donné par une matrice dont chaque vecteur représente le module de la TFD d'une trame du signal de parole :

$$SPEC = [|X_0||X_1|\dots|X_L|] \tag{I.2}$$

où L est le nombre de fenêtres du signal de parole. Le spectrogramme du signal de la figure I.11 est donné sur la figure I.12.

La taille de la fenêtre d'analyse est un paramètre important pour cette représentation. Pour de petites fenêtres (typiquement de l'ordre de 3 à 10 ms), on obtiendra une représentation avec une très bonne localisation temporelle mais avec une précision fréquentielle moins précise. On aura dans ce cas un spectrogramme à bande large. Dans le cas contraire où l'on choisit des fenêtres d'analyse de plus grande taille (typiquement supérieures à 20 ms), on obtient une plus grande précision fréquentielle au prix d'une localisation temporelle plus approximative. On parlera dans ce cas de spectrogramme à bande étroite. Pour la parole, les deux types de représentations sont utilisées suivant que l'on souhaite observer la structure fine du contenu fréquentiel (qui est clairement visible sur le spectrogramme à bande étroite) ou que l'on souhaite observer l'enveloppe

spectrale ou les formants (qui sont plus clairement visible sur un spectrogramme à bande large). La figure I.13 propose les spectrogrammes à bande étroite et à bande large d'une voyelle /a/ prononcée avec une fréquence fondamentale augmentant avec le temps. Les harmoniques sont alors très clairement identifiées sur le spectrogramme à bande étroite.

Les *formants* sont plus particulièrement visibles sur les spectrogrammes à large bande : ils sont matérialisés par des zones plus sombres indiquant des zones fréquentielles de plus forte énergie. Ils jouent un rôle important en parole et l'on peut déjà s'en rendre compte en observant le spectrogramme du signal /aeiou/ donné sur la figure I.14.

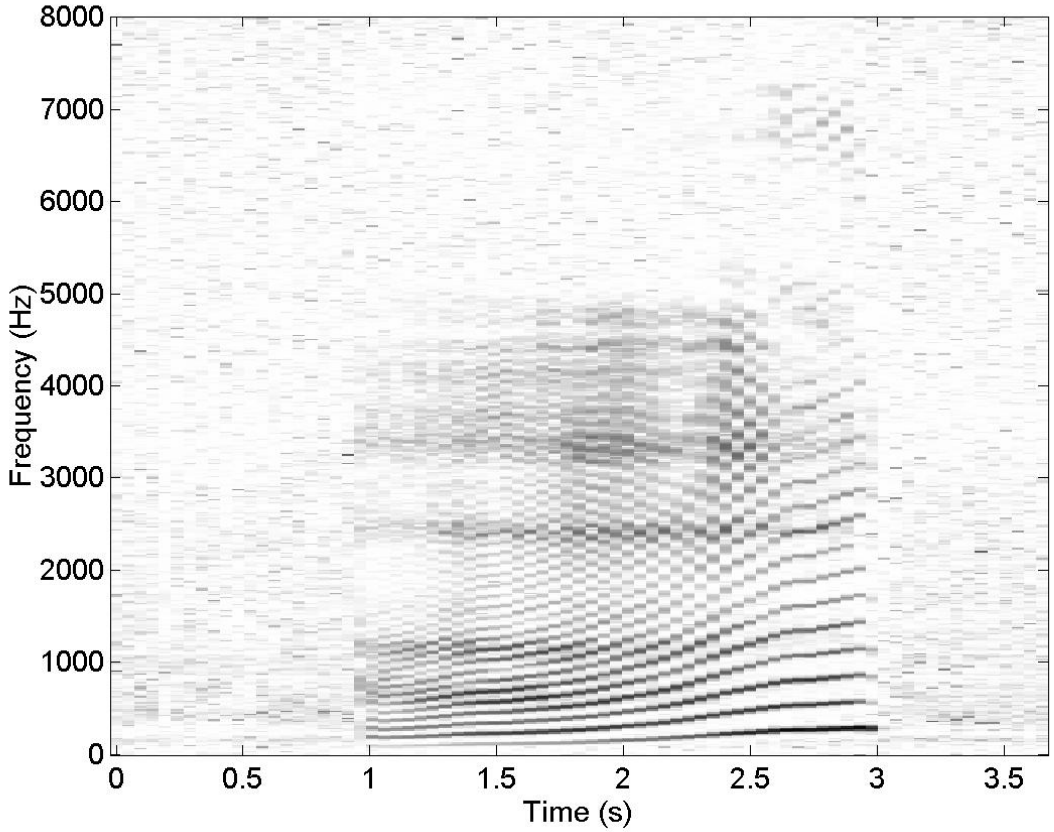
Sur ce spectrogramme, les mouvements brusques de ces formants, notamment les deux premiers, indiquent un changement de voyelle. Comme on le verra plus tard, on peut en effet caractériser les voyelles par la position de leurs seuls deux premiers formants. On ne tient pas compte en général du pic de très basse fréquence (autour de 200-300 Hz), parfois appelé formant glottal qui apparaît pour certaines voyelles ouvertes (notamment /a/ ou /ɛ/).

La figure I.15 représente le module de la TFD pour une trame du signal de parole (voyelle /i/). Cette représentation donne une "section" du spectrogramme et permet également de voir la structure fine (les harmoniques) et les formants à travers l'enveloppe spectrale.

Il est ainsi possible de représenter les voyelles en fonction de la position de leurs deux premiers formants F1 et F2. Cette représentation met en évidence une disposition en forme de triangle : on parle de *triangle vocalique*. On peut associer ce triangle vocalique au triangle articulatoire en reliant (de façon grossière) la position moyenne de la langue dans la cavité bucale : une position antérieure indique que la langue est proche des dents, une position postérieure que la langue est en arrière du conduit vocal, ouvert (resp. fermé) indiquant une position éloignée du palais (resp. près du palais donnant lieu à une constriction plus étroite , voir figure I.16)

Bien sur, en pratique, une voyelle suivant les locuteurs et suivant leur prononciation ne possédera pas une position des formants rigoureusement stable. La figure I.17 donne la position des deux premiers formants pour un nombre élevé d'élocutions de plusieurs voyelles par différentes personnes. Les ellipses représentent les régions grossières dans lesquelles on trouve la plus grande partie des occurrences de chaque voyelle.

On donne dans les figures suivantes un certain nombre de spectrogrammes permettant de mettre en évidence certaines caractéristiques des consonnes du français. Nous ne rentrerons pas ici dans le détail. On notera cependant la nature aléatoire (ou stochastique) du contenu fréquentiel des fricatives et la barre d'explosion caractéristique des plosives. On remarquera également que ces sons quoique moins énergétiques que les voyelles sont très étendus en fréquence.



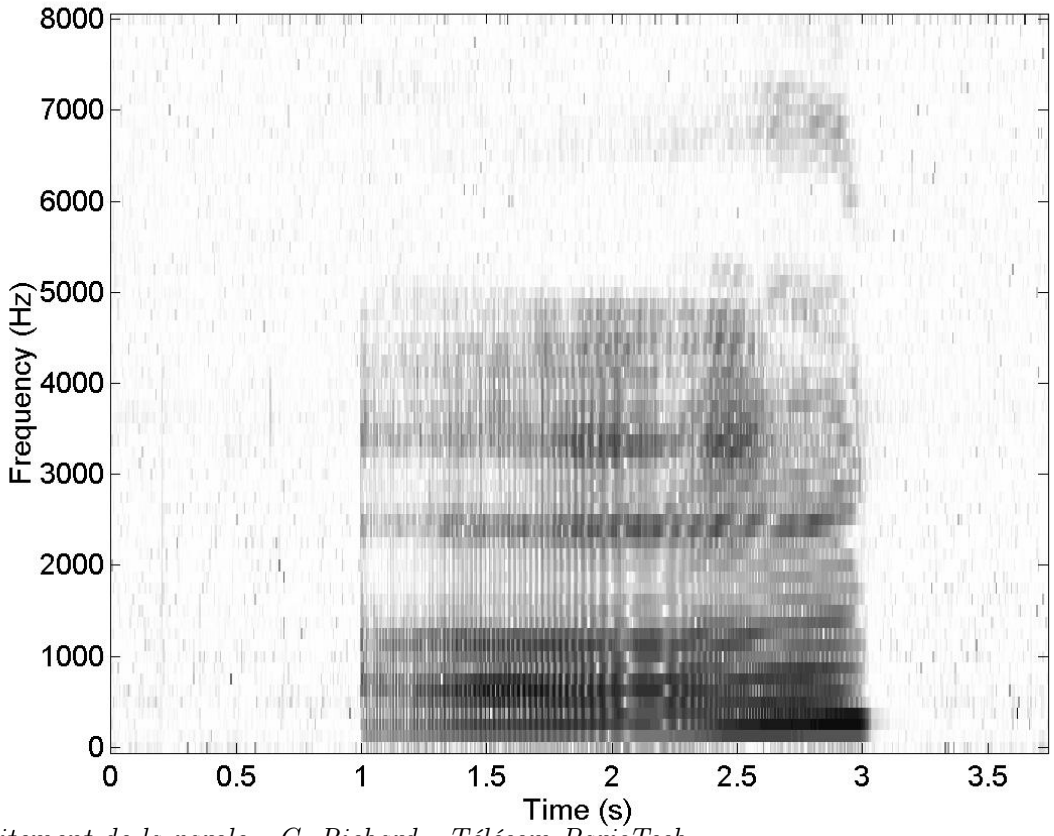


FIGURE I.13 – Spectrogramme bande étroite (haut) et spectrogramme large bande (bas) d'une voyelle /a/ produite avec une élevation progressive de la fréquence fondamentale"

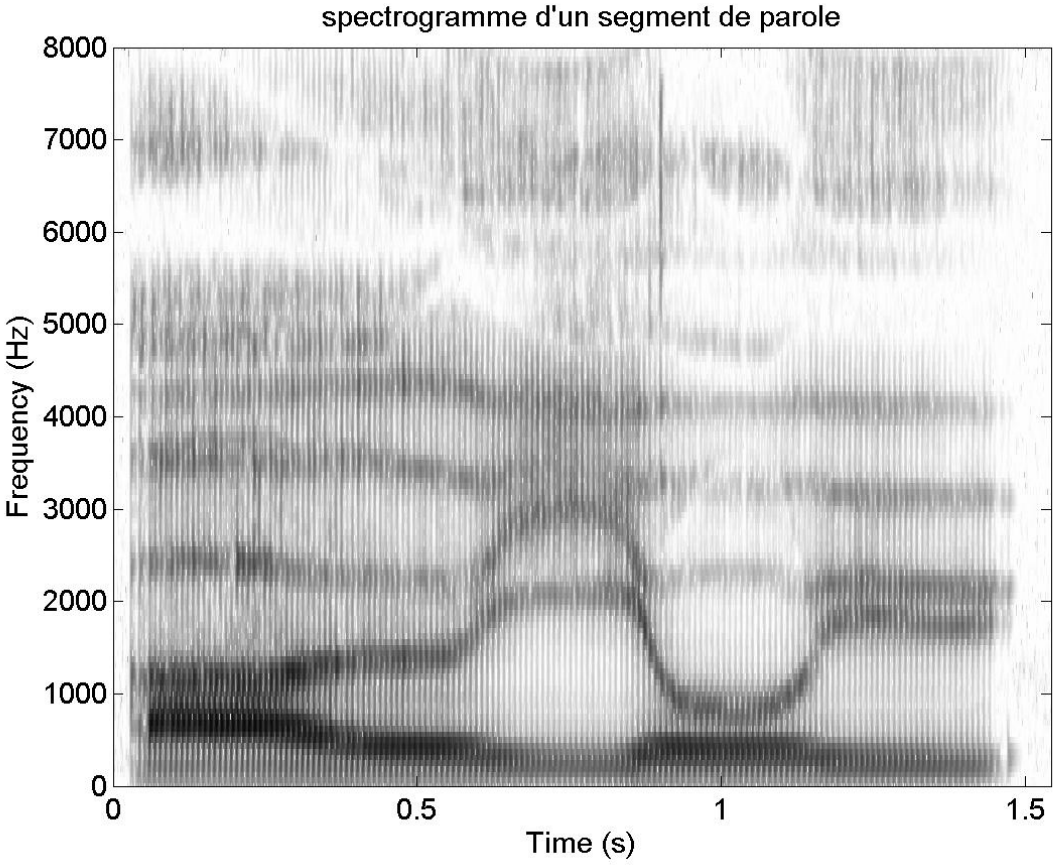


FIGURE I.14 – Spectrogramme du son constitué des voyelles /aeiou/ : les mouvements brusques des deux premiers formants, indiquent un changement de voyelle

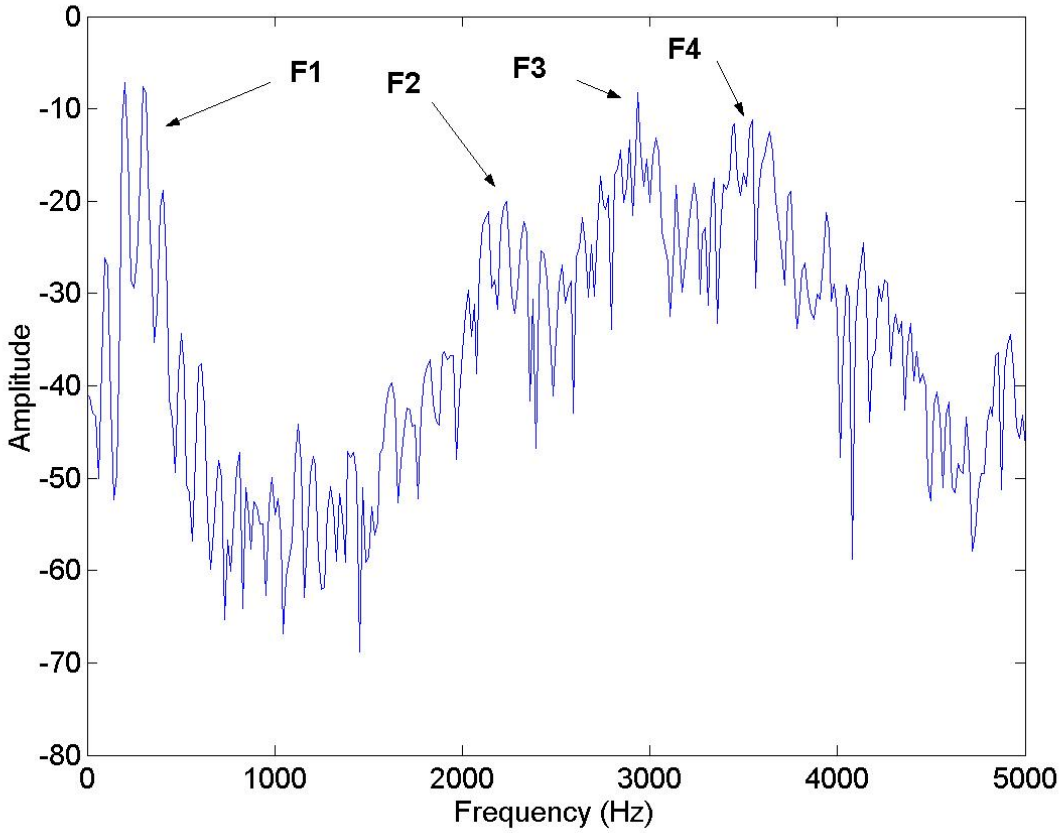


FIGURE I.15 – Module de la transformée de Fourrier (ou coupe spectrographique) d'une trame de la voyelle /i/

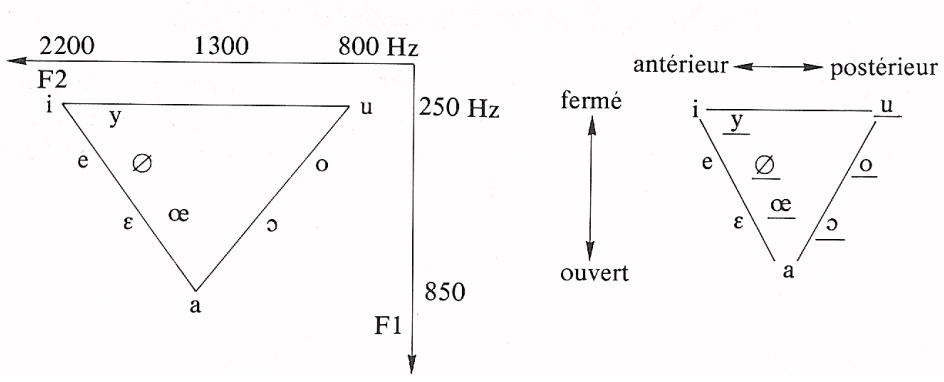


FIGURE I.16 – Triangle vocalique

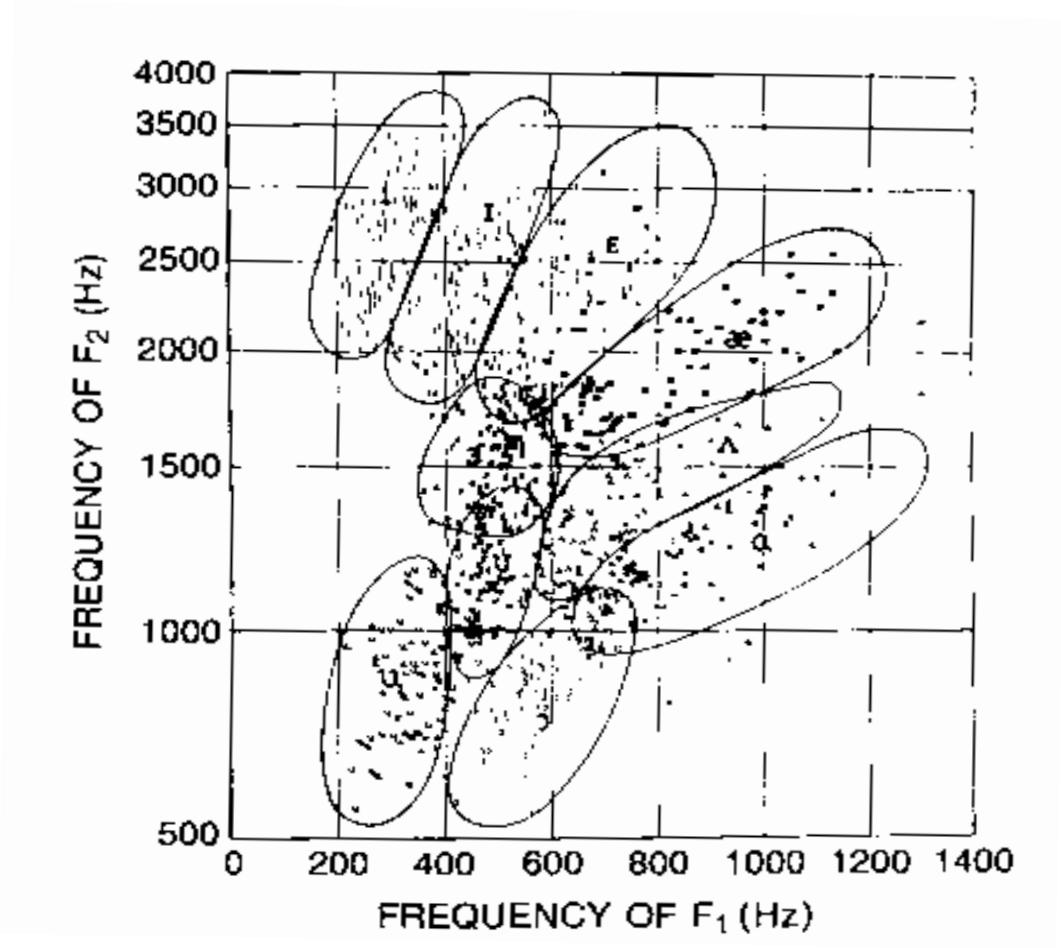


FIGURE I.17 – Représentation des sons vocaliques de l'anglais en fonctions des deux premières fréquences formantiques (D'après [24])

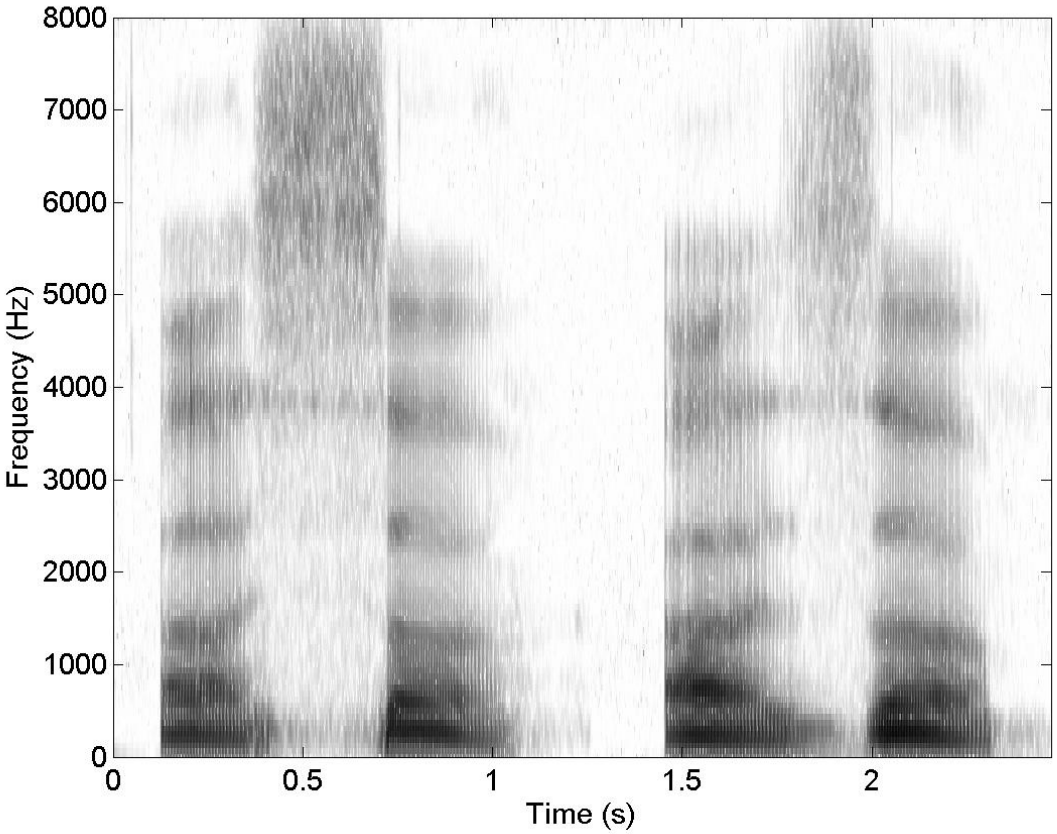


FIGURE I.18 – Spectrogramme bande large du signal "assa aza" (*/a s a a z a/*)"

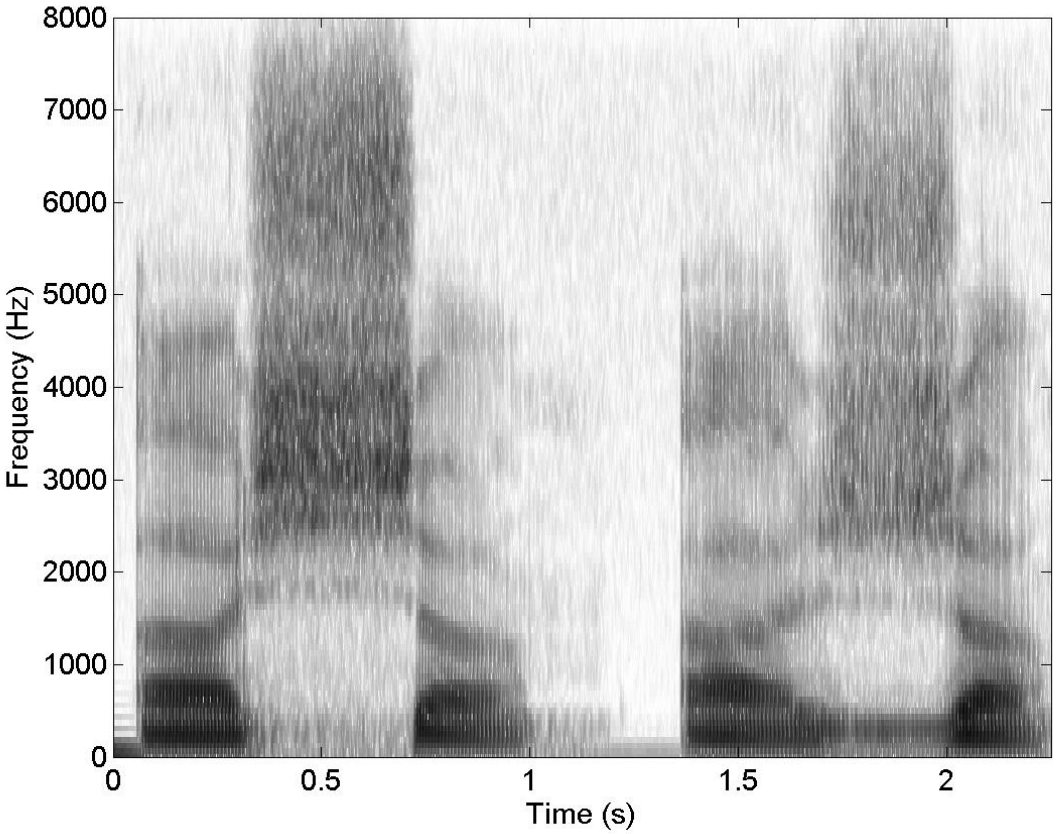


FIGURE I.19 – Spectrogramme bande large du signal "acha aja" (*/a ʃa a ʒa/*)"

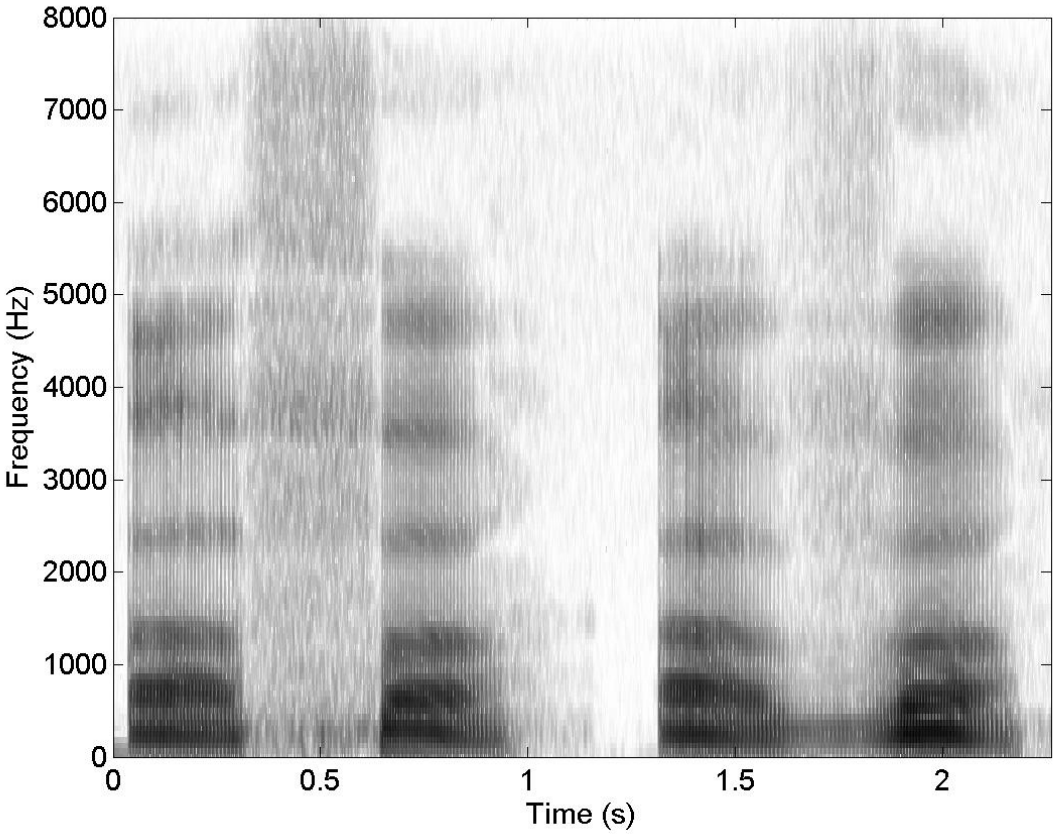


FIGURE I.20 – Spectrogramme bande large du signal "afa ava" (*/a f a a v a/*)"

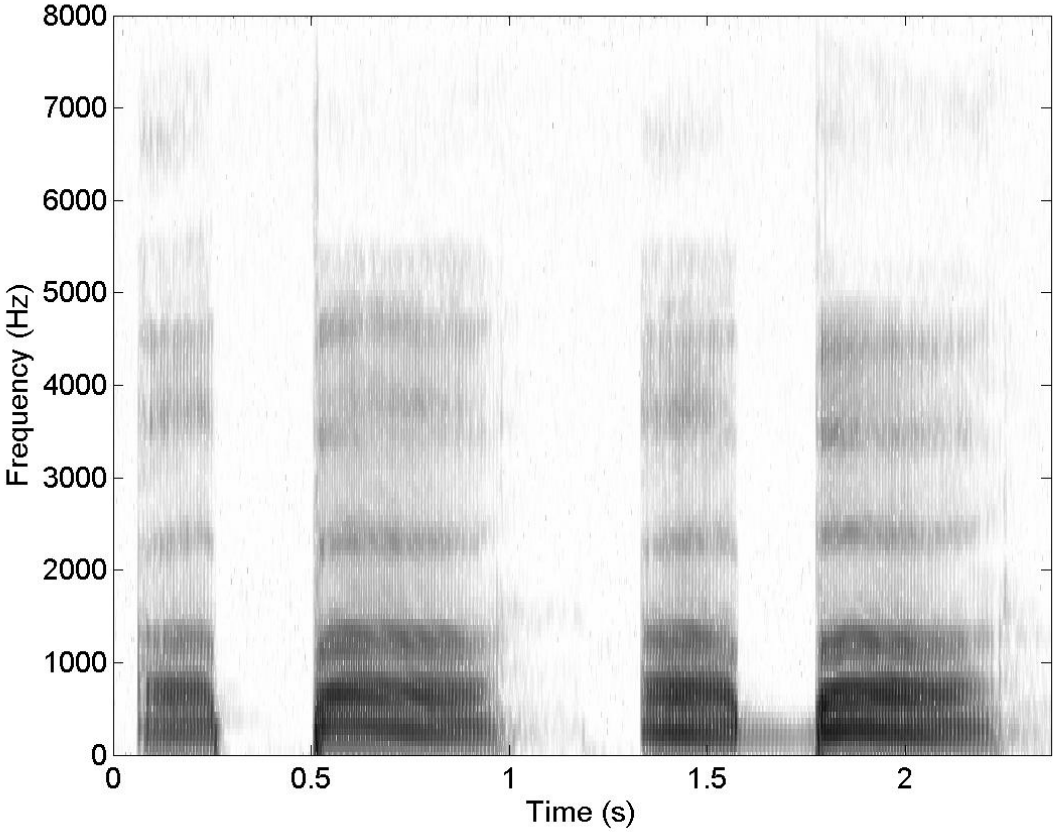


FIGURE I.21 – Spectrogramme bande large du signal "apa aba" (*/a p a a b a/*)"

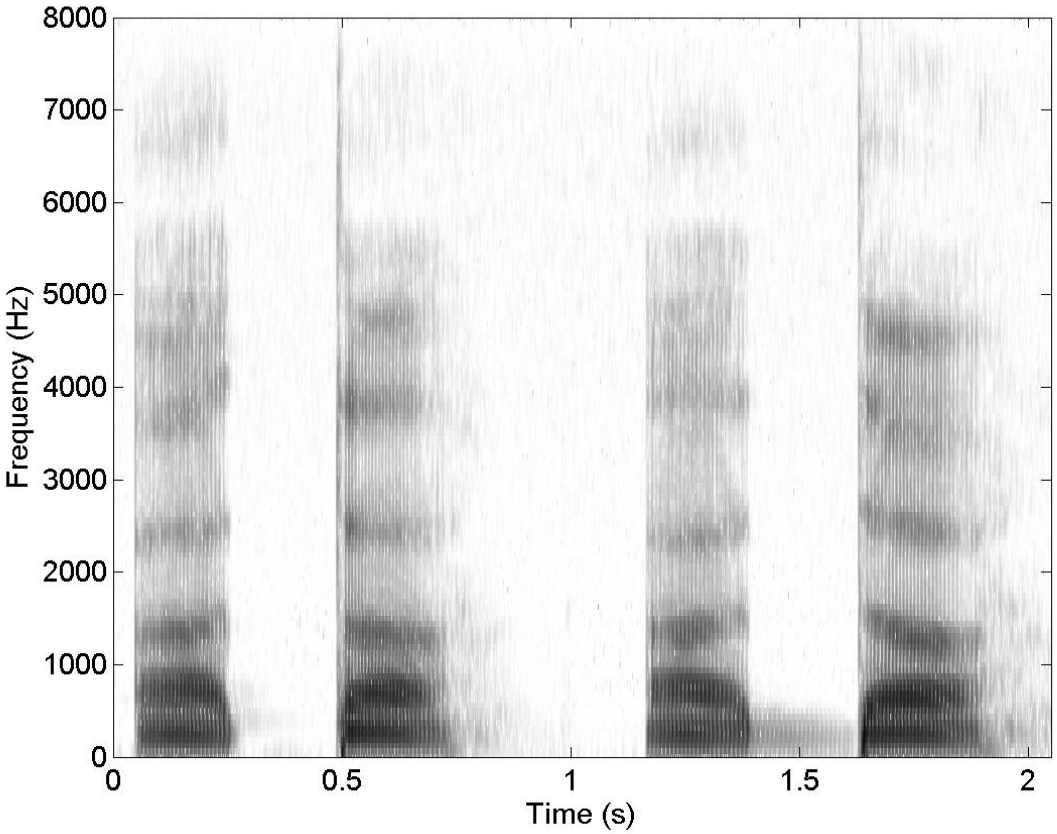


FIGURE I.22 – Spectrogramme bande large du signal "ata ada" (*/a t a a d a/*)"

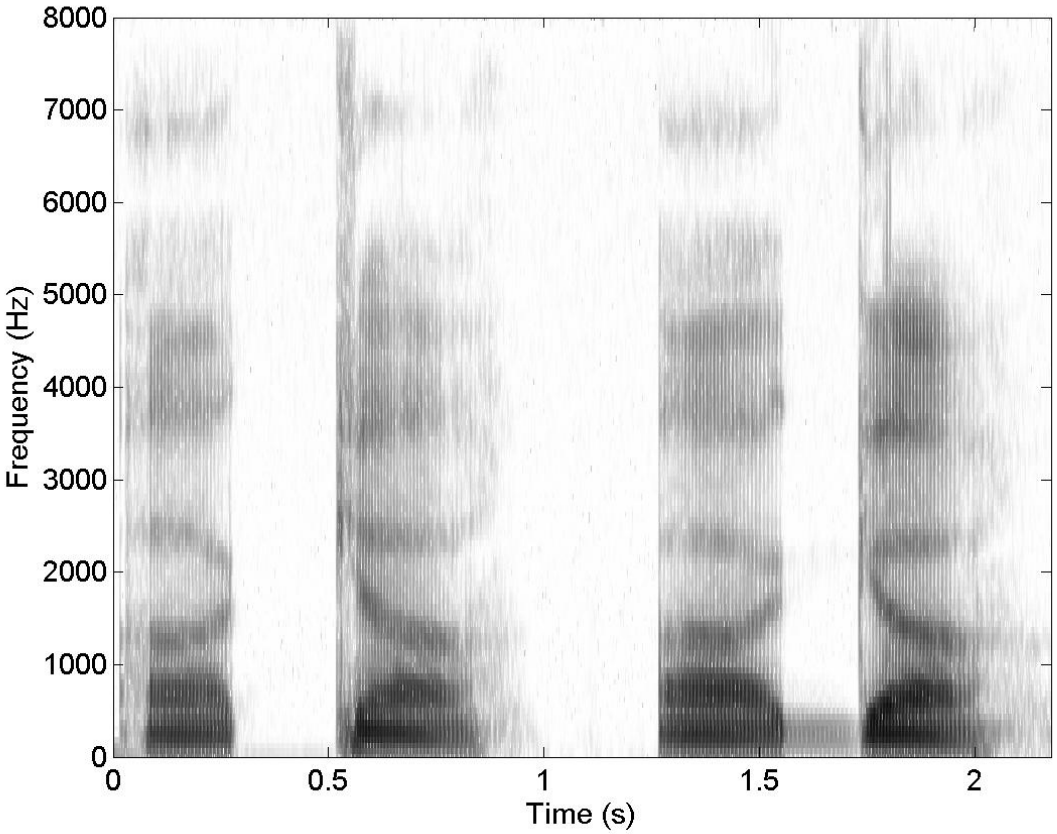


FIGURE I.23 – Spectrogramme bande large du signal "aka aga" (*/a k a a g a/*)"

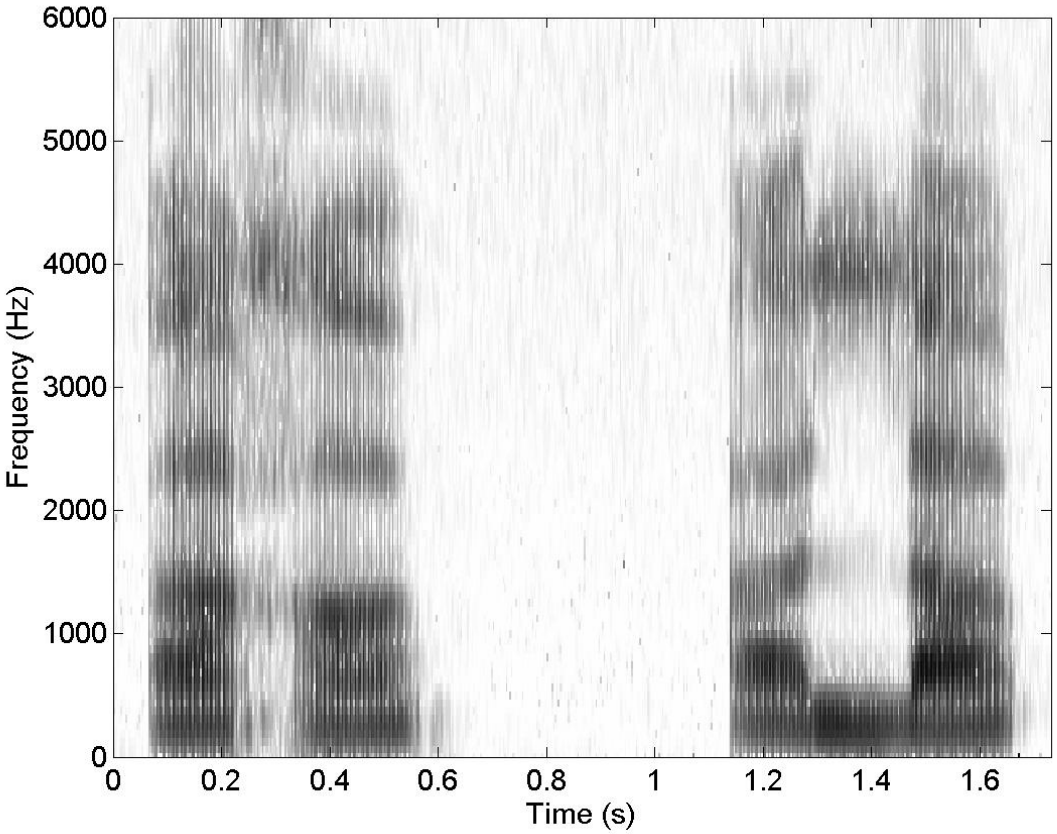


FIGURE I.24 – Spectrogramme bande large du signal "ara ala" (*/a R a a l a/*)"

Chapitre III

Synthèse de la parole [1]

III.1 Définition

Un système de synthèse à partir du texte (TTS : Text-To-Speech) est une machine capable de lire a priori n'importe quel texte à voix haute. Un tel système diffère fondamentalement d'autres machines parlantes en ceci qu'il est destiné à lire à voix haute des phrases qui n'ont en principe jamais été lues auparavant. Il est en effet possible de produire automatiquement de la parole en concaténant simplement des mots ou des parties de phrases préalablement enregistrées, mais il est clair dans ce cas que le vocabulaire utilisé doit rester très limité et que les phrases à produire doivent respecter une structure fixe, afin de maintenir dans des limites raisonnables la quantité de mémoire nécessaire à stocker les éléments vocaux de base. C'est le cas, par exemple, de l'horloge parlante ou les nombres appropriés sont insérés dans la phrase porteuse "au quatrième top il sera exactement (*nombre inséré*) heure (*nombre inséré*) minutes (*nombre inséré*) secondes". On définira donc plutôt la synthèse TTS comme la production automatique de phrases par calcul de leur transcription phonétique.

III.2 Architecture d'un système TTS

Comme on la vu précédemment, la parole naturelle est intrinsèquement soumise aux équations aux dérivées partielles de la mécanique des fluides, soumises de surcroît à des conditions dynamiques étant donné que la configuration de nos muscles articulateurs évolue dans le temps. Ceux-ci sont contrôlés par notre cortex, qui met à profit son architecture parallèle pour extraire l'essence du texte à lire : son sens. Même s'il semble aujourd'hui envisageable de construire un synthétiseur basé sur ces modèles, une telle machine présenterait un niveau de complexité peu compatible avec des critères économiques, et d'ailleurs probablement inutile. Il ne faut dès lors pas s'étonner si le fonctionnement interne des systèmes TTS développé à ce jour s'écarte souvent de leurs homologues humains.

La figure III.1 donne un schéma d'une architecture classique d'un système TTS. Un système TTS est en fait contitué de deux principaux blocs : *l'analyse du texte* et *la synthése* à proprement parlé.

– *L'analyse du texte* va fournir, à partir du texte initial, une transcription phonétique associée à des informations d'intonation et de rythme. Elle inclut les étapes de prétraitement du texte, de transcription graphème-phonèmes, et le module prosodique.

1. Chapitre reprenant de larges extraits du polycopié de cours de T. Dutoit [11] et du cours de F. Beaugendre [3]

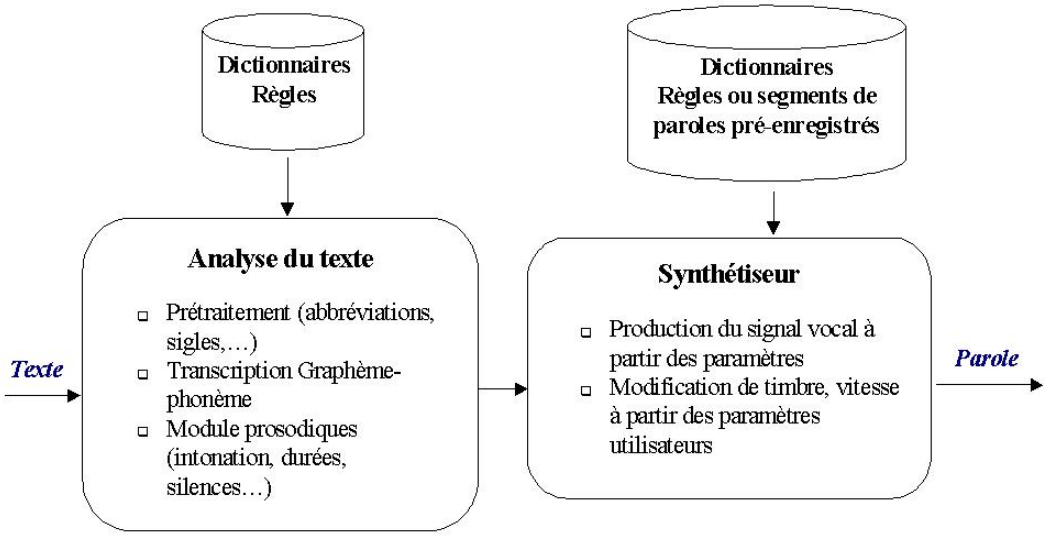


FIGURE III.1 – Architecture classique d'un système TTS

– La synthèse va produire le signal vocal à partir de la transcription phonétique et des informations prosodiques précédemment obtenues.

III.3 L'analyse du texte

L'organisation générale du module de traitement du texte (souvent dénommé module de traitement du langage naturel) est donnée à la figure III.2. On retrouve dans ce schéma un nombre conséquent de modules qui vont permettre d'obtenir une phonétisation du texte initial. Dès maintenant, nous pouvons remarquer les modules morphologique et syntaxique qui jouent un rôle prépondérant dans la phonétisation et qui sont décrits ci-dessous.

III.3.1 Le prétraitement du texte

Ce module de prétraitement a généralement deux rôles principaux. Le premier est un rôle d'interface entre le texte (représentation linéaire) et la structure de donnés internes gérée par le synthétiseur ([11]). Le second est un rôle d'identification des séquences de caractères qui risquent de poser un problème de prononciation. Parmi les problèmes principalement rencontrés, on peut citer :

– **La détection de fin de phrase** (localisation du point). Ce problème peut s'avérer délicat car le point peut être utilisé dans les nombres (nombres rationnels 3.14), dans les dates (3.3.2001), dans les abbréviations (resp.), dans les acronymes (E.N.S.T), ...
– **Le traitement des abbréviations**. Notons qu'il n'existe pas toujours une transcription unique pour une abbréviation donnée comme on peut le voir sur l'example suivant ([11])) : *"Dr. Jones lives at the corner of Jones Dr. And St. James St."*
– **Le traitement des acronymes** et notamment identifier si l'acronyme se prononce (comme pour OVNI ou CNET) ou s'épelle (comme pour SNCF). Cette décision n'est toujours immédiate car pourquoi prononce-t-on CNET et pas ENST ?
– **Le traitement des nombres** est un problème à part entière. Il s'agit notamment de pouvoir désambiguiser les nombres rationnels 2.05 des dates 2.05 (2 mai), des heures 2.05

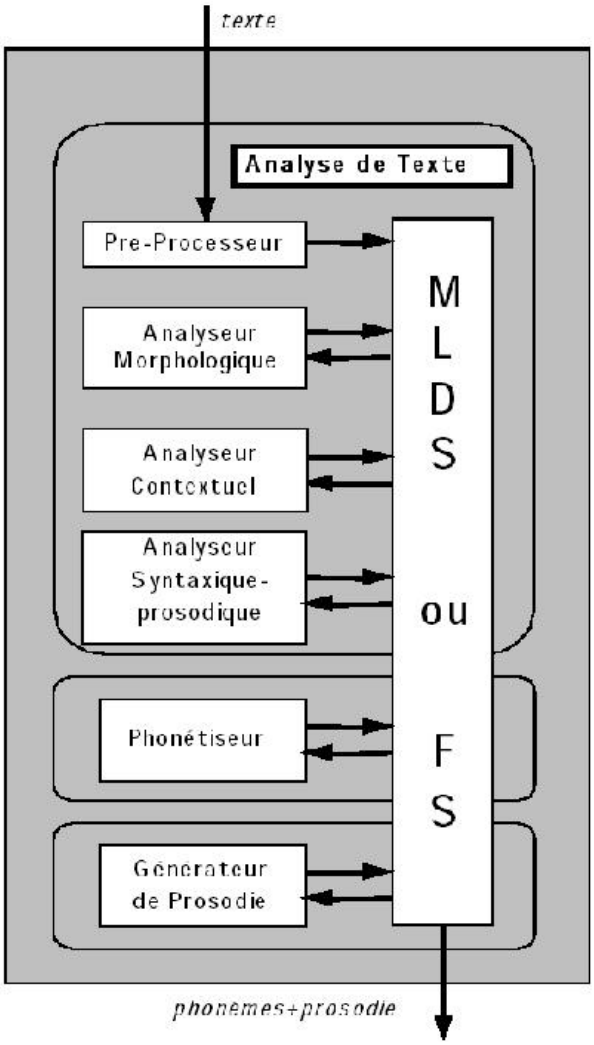


FIGURE III.2 – Module de traitement de texte (d'après [5])

(2h et 5 minutes), d'interpréter les chiffres romains (Henri IV) etc.
Le problème du pré-traitement du texte est particulièrement important pour les applications de lecture de Méls ou de pages Web. En effet, dans de tels textes on trouve fréquemment des abbrévations et sigles mais également des smileys (-) ;..), des adresses WEB (http ://tsi.enst.fr/g̃richard/), des adresses mél (gael.richard@enst.fr).

III.3.2 Analyse morphologique

Un analyseur morphologique a pour but de proposer toutes les natures possibles pour chaque mot pris individuellement, en fonction de sa graphie. Les intérêts de l'analyse morphologique sont les suivants :

- Elle permet de réduire la taille des lexiques/dictionnaires
- Elle permet d'obtenir des informations sur la catégorie syntaxique des mots et a ainsi une grande influence sur la prosodie
- Elle permet ainsi d'aider la traduction graphème-phonème (C'est l'analyse morphologique qui permettra de savoir que le /s/ de présupposer de ne se prononce pas /z/)
- Pour certaines langues (allemand), elle permet de prédire la position de l'accent. En allemand l'accent tombe souvent sur la première syllabe de la racine d'un mot (ex : 'Band, Ver'Band)

Cette analyse distingue en général deux catégories de mots :

- *Les mots grammaticaux* qui forment le squelette syntaxique de la phrase. Ce sont les déterminants, les pronoms, les prépositions et les conjontions. Les mots grammaticaux sont en nombre finis (moins de 1000 en français). Ils sont en général mémorisés dans un lexique qui associe leur graphie à leur prononciation (et à leur nature grammaticale).
- *Les mots lexicaux* qui sont a priori en nombre infini. Ils nécessiteront ainsi une analyse morphologique plus poussée (morphologie inflectionnelle, dérivationnelle et compositionnelle).

Notons que l'analyse morphologique décrit les mots d'une langue en termes d'unités élémentaires de sens qui sont appelées les morphèmes. Ces unités abstraites peuvent aussi bien représenter un mot qui a du sens qu'un concept grammatical (morphème pluriel). Ainsi "fasse" combine le morphème "faire" avec le morphème du conditionnel présent (3ième personne du singulier). Notons, enfin qu'un morphème peut très bien ne pas apparaître comme c'est le cas dans "des choix" ou le morphème pluriel n'a pas de correspondance graphémique.

Analyse inflexionnelle

L'importance de l'analyse inflectionnelle n'est pas la même pour toutes les langues. L'anglais par exemple fait un usage très réduit de l'inflexion ce qui n'est pas le cas du français. Comme son nom l'indique l'analyse inflexionnelle va chercher à déterminer les inflexions possibles à partir de formes de base (encore appelés *lexèmes* ou racines). Un exemple d'analyse inflexionnelle est donné ci-dessous à travers une approche déclarative :

$$\begin{cases} \textbf{Lexème :} & \text{groupe\_d\_inflexion, racine\_1, } \ldots \text{, racine\_N,} \\ \textbf{groupe\_d\_inflexion :} & \text{mode\_d\_inflexion\_1, groupe\_de \_suffixe\_1, i,j, } \ldots \text{,k.} \\ & \text{mode\_d\_inflexion\_2, groupe\_de\_suffixe\_2, l,m, } \ldots \text{,n.} \\ & \ldots . \\ \textbf{Groupe\_de\_suffixe\_1 :} & \text{suffixe\_11, } \ldots \text{, suffixe\_1N,.} \end{cases}$$

On peut par exemple donner l'exemple suivant tiré du verbe tenir :

$$\begin{cases} \textit{tenir}: & \text{venir, tien, ten,tienn,tin, tîn,} \\ \textit{venir}: & \text{indicatif\_présent, suf\_ind\_prés, 1, 1, 1, 2, 2, 3} \\ & \text{subjonctif\_présent, suf\_subj\_prés, 3, 3, 3, 2, 2, 3} \\ & \text{etc....} \end{cases}$$

ce qui donne les formes suivantes :

- je tiens, tu tiens, il tient, nous tenons, vous tenez, ils tiennent
- Que je tienne, que tu tiennes, que nous tenions, que vous teniez, qu'ils tiennent
- etc ...

Analyse dérivationnelle et compositionnelle

L'analyse dérivationnelle va comme pour l'analyse inflexionnelle s'intéresser aux transformations morphologiques que peut subir une forme de base (léxème). Dans l'analyse dérivationnelle, on s'intérèsse aux différentes dérivations qui peuvent être obtenues à partir d'une forme de base. On peut ainsi dire que c'est l'étude de la construction des mots de catégories syntaxiques différentes à partir d'un morphème de base. Par exemple, à partir du morphème de base *image*, on peut en déduire : *imagine*, *imagination*, ou encore *imagerie*.

L'analyse compositionnelle est en fait l'étude de la composition de mots à partir de plusieurs morphèmes de base (par exemple Porte + avion = porte-avion) . Dans certaines langues (comme l'allemand), cette étude peut s'avérer particulièrement complexe (voir [5] pour un exemple). En français, ce problème est toutefois moins délicat.

III.3.3 Analyse contextuelle et syntaxique

Un analyseur contextuel, qui considère les mots dans leur contexte, à pour but
- de réduire le nombre de catégories possibles pour chaque mot en fonction de ses voisins
- de fournir un étiquetage de chacune des unités lexicales constituant la phrase.

On pourrait croire que chaque mot est rattaché à une classe possible, ce qui est loin d'être le cas. Par exemple en français, *le* peut être pronom ou déterminant, *joue* peut être un verbe ou un nom et *couvent*, peut être un nom ou un verbe (dans ce dernier cas, on remarquera d'ailleurs qu'une erreur de classification impliquera une erreur de transcription graphème-phonème).

Il existe un grand nombre d'approches pour l'analyse syntaxique contextuelle. On pourra consulter [5] pour obtenir un premier panorama de ces méthodes. De façon plus succincte, notons qu'il y a 2 principaux types d'approches :
- *Les analyseurs déterministes* qui exploitent un ensemble de règles catégoriques (par exemple de type oui/non). Cette approche est en général assez complexe et nécessite une grande expertise de la langue.
- *Les analyseurs probabilistes* qui utilisent les probabilités de transitions entre catégories syntaxiques successives. Ces approches, assez simples, sont en général très performantes mais nécessitent d'avoir d'importants corpus de données. Nous donnons un exemple ci-dessous à travers les modèles n-grammes.

Etiquetage par n-grammes

Les n-grammes sont utilisés pour estimer la probablité d'une suite de mots $w_1, w_2, ..., w_n$ pour un langage donné (notons ici que langage peut représenter une langue mais aussi plus précisément l'utilisation d'une langue dans un domaine d'application particulier). Les modèles

les plus répandus sont les *modèles bigrammes* (la probabilité d'un mot ne dépend que de celle du mot précédent) et les *modèles trigrammes* (la probabilité d'un mot ne dépend que de celle des 2 mots précédents). Ainsi, on peut écrire le modèle trigramme d'une suite de mot $w_1, w_2, ..., w_n$ sous la forme :

$$P(w_1, w_2, \ldots, w_N) \approx P(w_1)P(w_2|w_1)P(w_3|w_2, w_1)\ldots P(w_N|W_{N-1}, W_{N-2}) \quad \text{(III.1)}$$

Pour la synthèse de parole, c'est principalement la suite des étiquettes syntaxiques $\hat{\mathbf{T}}$ que l'on cherche à obtenir parmir toutes les suites d'étiquettes $\mathbf{T} = (t^1, t^2, ...t^N)$ possibles où t^i est l'étiquette associée au mot w_i. Cela peut s'écrire :

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{W}) \quad \text{(III.2)}$$

En appliquant la règle de Bayes, on obtient :

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}} \frac{P(\mathbf{T}|\mathbf{W})}{P(\mathbf{W})} = arg\max_{\mathbf{T}} \frac{P(\mathbf{W}|\mathbf{T})P(\mathbf{T})}{P(\mathbf{W})} \quad \text{(III.3)}$$

Pour effectuer l'étiquetage de la suite de mots, on est amené a de plus supposer que :
- La probabilité d'un mot étant donné le passé ne dépend que de son étiquette (hypothèse restreignant l'application du modèle à l'étiquetage syntaxique)
- Le probabiité d'une étiquette étant donné le passé ne dépend que des 2 étiquettes précédentes (modèle trigramme).

Nous pouvons alors écrire que :

$$P(\mathbf{W}|\mathbf{T}) = P(w_1, w_2, ..., w_N|t^1, t^2, ..., t^N) \approx \prod_{i=1}^{N} P(w_i|t^i) \quad \text{(III.4)}$$

$$P(\mathbf{T}) = P(t^1, t^2, ..., t^N) \approx \prod_{i=1}^{N} P(t_i|t^{i-1}, t^{i-2}) \quad \text{(III.5)}$$

Le modèle trigramme de l'étiquetage syntaxique par trigrammes se résume ainsi à :

$$P(t^1, t^2, ..., t^N|w_1, w_2, ..., w_N) \approx \prod_{i=1}^{N} P(w_i|t^i)P(t_i|t^{i-1}, t^{i-2}) \quad \text{(III.6)}$$

On associe alors un automate probabiliste à l'équation III.6. Par exemple, pour un modèle bigramme, le modèle comporte M états où chaque état q_i est associé à une étiquette donné t^i. Par souci de simplicité, on autorise toutes les transitions possible, et on utilisera ainsi un automate complètement connecté (encore appelé ergodique). Pour plus de précisions, on pourra se rapporter à [5].

Notons que les deux approches (déterministes et probabilistes) ont maintenant tendance à se rapprocher et on voit ainsi de nouvelles approches qui visent à obtenir une grammaire locale (déterministe) dont les règles sont obtenues par inférence automatique à partir de grands corpus.

Notons, pour conclure cette partie, que l'analyse syntaxique va permettre également d'aider l'analyseur syntaxique-prosodique, qui va établir un découpage du texte en groupes de mots, ce qui permettra d'y associer une prosodie.

III.3.4 Transcription graphème-phonème

Le but de cette transcription est de transformer un texte orthographique (graphème) sous forme d'un texte phonétique ou liste de phonèmes. On utilise pour cela un alphabet phonétique (voir chapitre II) qui spécifie les sons élémentaires des langues parlées. Effectuer une telle transcription est plus ou moins difficile suivant la langue et elle est en particulier plutôt complexe pour le français. En effet, pour cette langue on trouve une grande variété de prononciations pour une graphie donnée. Ainsi par exemple le x, le /ch/ ou le /s/ possèdent plusieurs prononciations possibles :

- **'x'** se prononce : [ks] dans le mot axe, [s] dans six, [z] dans sixième, [gz] dans exact,
- **'s'** se prononce : [z] dans le mot doser, [s] dans les mots parasol entresol, ou ne prononce pas du tout (pluriel)
- **'ch'** se prononce : [k] dans le mot chlore, [$\int$] dans château,
- ...

A l'opposé, des graphies différentes peuvent donner lieu à un phonème identique (ce qui crée évidemment moins de problèmes pour la synthèse !) :

- le phonème [/ε/] se retrouve dans les mots 'mère', 'fête', 'fer', 'peine', 'sept', 'aspect', 'est', 'relais', 'tramway','laid', 'monnaies'
- Le phonème [o] se retrouve dans les mots 'pot', 'peau', 'auréole'
- ...

De manière plus générale, il existe un certain nombre de phénomènes qui rendent la traduction graphème-phonème difficile. Ce sont :

Les homographes-hétérophones : Ce sont les mots qui s'orthographient de la même façon (homographes) mais qui se prononcent différemment (hétérophones). Quoique moins fréquents que les homographes-homophones, le français standard comprend environ 150 homographes-hétérophones. La plupart d'entre eux partagent une racine commune (par exemple : *un président* /ils *président* ; *somnolent* / ils *somnolent*) mais ce n'est pas toujours le cas (les *portions* / nous *portions* ; les *fils* à papa / les *fils* de nylon)

Les assimilations : elles sont principalement dues à la coarticulation où les contraintes articulatoires induisent des changements de prononciation. Ce phénomène peut générer d'importantes sources de variation phonétique (par exemple *'Absent'* sera prononcé *'apsent'*). On peut également observer un phénomène appelé *harmonisation vocalique* qui peut ouvrir une voyelle originellement fermée (par exemple /e/ devient /ε/ dans les mots *céderait*, *événement*).

Les liaisons : elles se caractérisent par l'inclusion d'un phonème à la frontière de deux mots. Ce phénomène est un cas particulier du français. Le nombre de liaisons effectuées dépend du niveau de langue et du style de prononciation. Certaines liaisons sont obligatoires et sont donc toujours faites (par exemple *'Très utiles'*), dautres sont optionnelles (par exemple *'Deux à deux'*), dautres encore sont interdites (par exemple *'Plat exquis'*). La présence d'une liaison dépend souvent des classes syntaxiques des mots concernés.

Le "e" muet : représente un problème plus complexe qu'il n'y parait. Rappelons que le "e" muet (ou schwa) est le phonème terminal que l'on trouve par exemple dans le mot 'table'. L'une des règles les plus courantes pour la prononciation (ou non) du "e" muet est celle des 3 consonnes : *"Un «e» est prononcé si sa disparition provoque le rapprochement de 3 consonnes"* (par ex : *table rouge*). En pratique, le problème s'avère plus complexe et il faudra tenir compte de contraintes rythmiques (le "e" muet est souvent prononcé en début de groupe rythmique comme dans *"pesez-les"*).

Noms propres, noms de lieu, nouveaux mots Pour ces mots, il est parfois nécessaire d'utiliser des règles phonologiques différentes de celles du français standard (par exemple pour des mots tels que *Schiltigheim, Ploumanach Reagan, Lendl, Pierce, Washington,* etc). Il peut être nécessaire d'essayer de détecter la langue source (par exemple pour *handball, football, revolver*). Enfin pour les nouveaux mots, l'approche couramment retenue consiste à utiliser des racines connues à partir desquelles il est possible de dériver ces nouveaux mots

Il est clair que ce sont les différentes analyses du texte décrites plus haut qui vont aider à obtenir une phonétisation automatique du texte.

De façon générale, il existe deux types d'approches pour la phonétisation automatique :

L'approche par dictionnaire : où le maximum de connaissances morphologiques sont concentrées dans un lexique. Parfois, on utilise des règles morphologiques pour déduire à partir des racines morphologiques stockées dans le lexique, les formes fléchies par dérivation, inflexion ou composition. Dans cette approche, seuls les mots non phonétisés par le dictionnaire sont alors transcrits par règles. C'est l'approche traditionnellement suivie pour l'anglais américain (MITALK) ([2])

L'approche par règles qui, à l'opposée de la précédente approche, utilise un maximum de règles pour décrire les connaissances phonologiques et n'utilise un lexique que pour phonétiser les exceptions. Il existe dans ce cadre un grand nombre de méthodes (incluant les systèmes experts, les méthodes avec apprentissage automatique des règles à partir d'une modélisation par chaînes de Markov ou neuronale). A ce jour, les approches les plus utilisées sont les approches "systèmes experts" qui se fondent sur des règles écrites par des experts (linguistes). La méthode la plus simple consiste à utiliser le contexte graphémique pour résoudre les conflits et a ainsi définir un ensemble de règles de réécriture sous la forme :

$$a \rightarrow [b]/l\_r : \mathrm{C} \tag{III.7}$$

qui se lit *"le segment a est réécrit en un segment b lorsquil est entouré des chaînes l et r (à gauche et à droite) et si la condition C est vérifiée"*. Nous donnons, ci-dessous, un exemple simple de fonctionnement d'une telle approche pour la phonétisation du mot "oiseau".

- Le mot "oiseau" se transcrit phonétiquement "/wazo/" , par application des règles suivantes :
  1. la chaîne de caractères orthographiques "oi" se transcrit par la succession des phonèmes /wa/, parce qu'elle est précédée d'un séparateur de mot et qu'elle n'est pas suivie de de la chaîne "gn" comme dans "oignon", ou d'un "n" comme dans "oindre".
  2. La lettre "s" se transcrit par le phonème /z/ car cette lettre est entourée par deux voyelles et que "oiseau" ne fait pas partie d'une liste d'exceptions à cette règle, stockée dans le lexique (on pense en particulier à "paraSol" ou "vraiSemblance").
  3. La chaîne de caractères "eau" se transcrit par le phonème /o/, indépendamment du contexte.

  De façon général, un système minimal en français nécessitera 500 règles, sachant qu'il faudra environ 1500 règles pour obtenir un système performant. Pour certaines langues (espagnol ou italien par exemple), d'excellentes performances peuvent être obtenues avec moins de 100 règles.

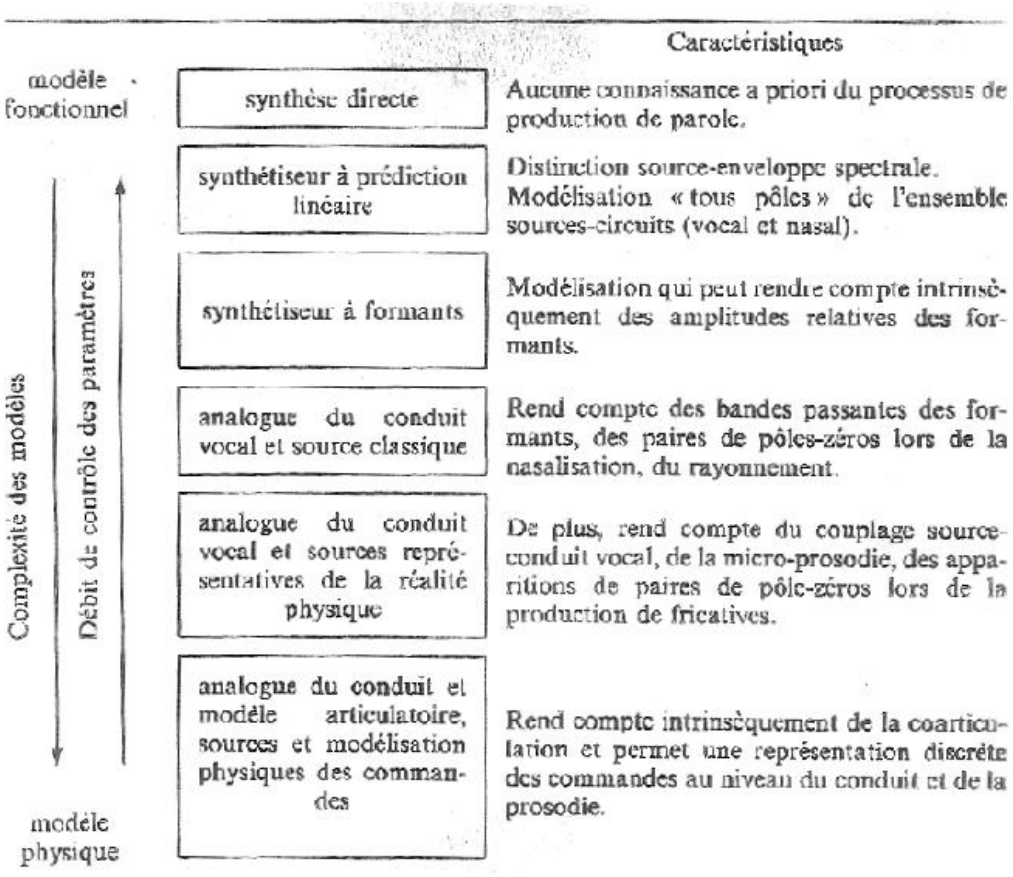


FIGURE III.3 – classification des techniques de synthèse (d'après [7])

III.4 Synthèse de la parole

L'étape précédente a permis de passer d'une information graphémique (le texte) en une information phonétique (les phonèmes à prononcer) associée à une information prosodique (l'intonation à reproduire). Cette dernière étape n'est pas décrite ici et n'est a priori pas essentielle pour un module simple de synthèse comme celui qui pourrait être développé dans PACT. L'étape de synthèse a proprement parlé consistera à générer le signal de parole à partir des informations pré-citées. Il existe de nombreuses approches pour réaliser cette synthèse. Ces approches sont classiquement ordonnées en fonction de leur niveau de description allant des modèles fonctionnels (synthèse par concaténation d'éléments sonores pré-enregistrés) aux modèles physiques (synthèse articulatoire). La figure III.3 donne une telle classification des techniques de synthèse en précisant les caractéristiques principales de chaque approche.

Nous ne décrirons ici que l'approche par concaténation.

III.4.1 Synthèse par concaténation

Au contraire des synthétiseurs par règles, les synthétiseurs par concaténation ont une connaissance très limitée du signal qu'ils mettent en forme. La plupart de ces connaissances se trouve en effet stockée dans les unités de parole mises en oeuvre par le synthétiseur. Ceci apparaît clairement dans la description générale d'un tel synthétiseur (voir figure III.4), où l'on constate

que la plupart des opérations liées à la synthèse proprement dite (par opposition aux opérations nécessaires à la création du synthétiseur) se retrouvent groupées dans un bloc de traitement du signal ne faisant aucune référence explicite à la nature profonde des signaux traités. La synthèse par concaténation procède en effet par mise bout à bout de segments acoustiques déjà coarticulés, extraits d'une base de données de signaux de parole. Il s'ensuit que, contrairement aux cibles phonétiques de l'approche précédente, qui nécessitent l'établissement de règles (phonétiques) pour modéliser correctement leurs transitions, la production de parole fluide en synthèse par concaténation ne requiert qu'une étape de concaténation qui s'accompagne d'un lissage purement acoustique des discontinuités pouvant apparaître aux points de concaténation. Comme pour la synthèse par règles, un certain nombre d'opérations préliminaires doivent être menées avant que le synthétiseur ne soit capable de produire sa première parole. C'est le rôle des modules de traitement de la parole.

Sélection des unités de synthèse

On commence ainsi par sélectionner les unités de parole qui devront permettre de minimiser les futurs problèmes de concaténation. Comme on le verra ci-dessous, cette sélection peut être soit statique (un seul choix possible par unité) soit dynamique (plusieurs choix possibles pour chaque unité, le choix étant fait au moment de la synthèse en fonctions de divers paramètres).

Diverses combinaisons *de diphones* (un diphone est une unité acoustique qui commence au milieu de la zone stable d'un phonème et se termine au milieu de la zone stable du phonème suivant), de *demi-syllabes*, et *de triphones* (qui diffèrent des diphones en ceci qu'ils comprennent un phonème central complet) sont en général retenues, dans la mesure où elles incluent assez correctement les phénomènes de coarticulation tout en ne nécessitant qu'un nombre limité d'unités. Dans le cas de phonèmes ne présentant pas de partie stationnaire, on prend soit la partie la plus stable, soit un triphone, ce qui évite de devoir segmenter dans une partie transitoire.

Constitution de la base de données de segments

On établit ensuite un corpus textuel (liste de mots, de courtes phrases, voire de textes) dans laquelle toutes les unités choisies apparaissent au moins une fois (plus si possible, de façon à ne pas devoir procéder à plusieurs enregistrements successifs si certaines des unités sont mal enregistrées). On peut dès à présent distinguer deux approches lors de la constitution de ce corpus.

Dans la première, que nous appellerons synthèse à sélection segmentale d'unités, on considère que toutes les instances d'une même unité phonétique sont équivalentes. Dans le cas d'une synthèse par diphones, par exemple, cela conduira à ne retenir qu'une version de chaque diphone et à s'arranger plus tard (lors de la synthèse proprement dite) pour en modifier la durée et/ou le pitch lors d'une étape dite de modification de prosodie.

Au contraire, dans une approche récente que nous qualifierons de synthèse à sélection totale d'unités (totale étant pris ici au sens de segmental et supra-segmental), les caractéristiques suprasegmentales des sons sont également prises en considération pour leur sélection dans la base de données. Si l'on reprend le cas d'une synthèse par diphones, on retiendra alors un grand nombre de versions de chaque diphone, différant entre elles par leur durée et leur pitch. L'étape de modification de prosodie mentionnée plus haut s'en trouvera donc considérablement simplifiée (mais non pas totalement éliminée, puisqu'il est en principe impossible d'enregistrer un corpus reprenant toutes les durées et toutes les courbes mélodiques possibles pour chaque unité). On enregistre alors ce corpus sous forme numérique et on le segmente en unités, soit à la main, par inspection du signal à l'aide d'outils de visualisation (de spectrogrammes, principalement), soit

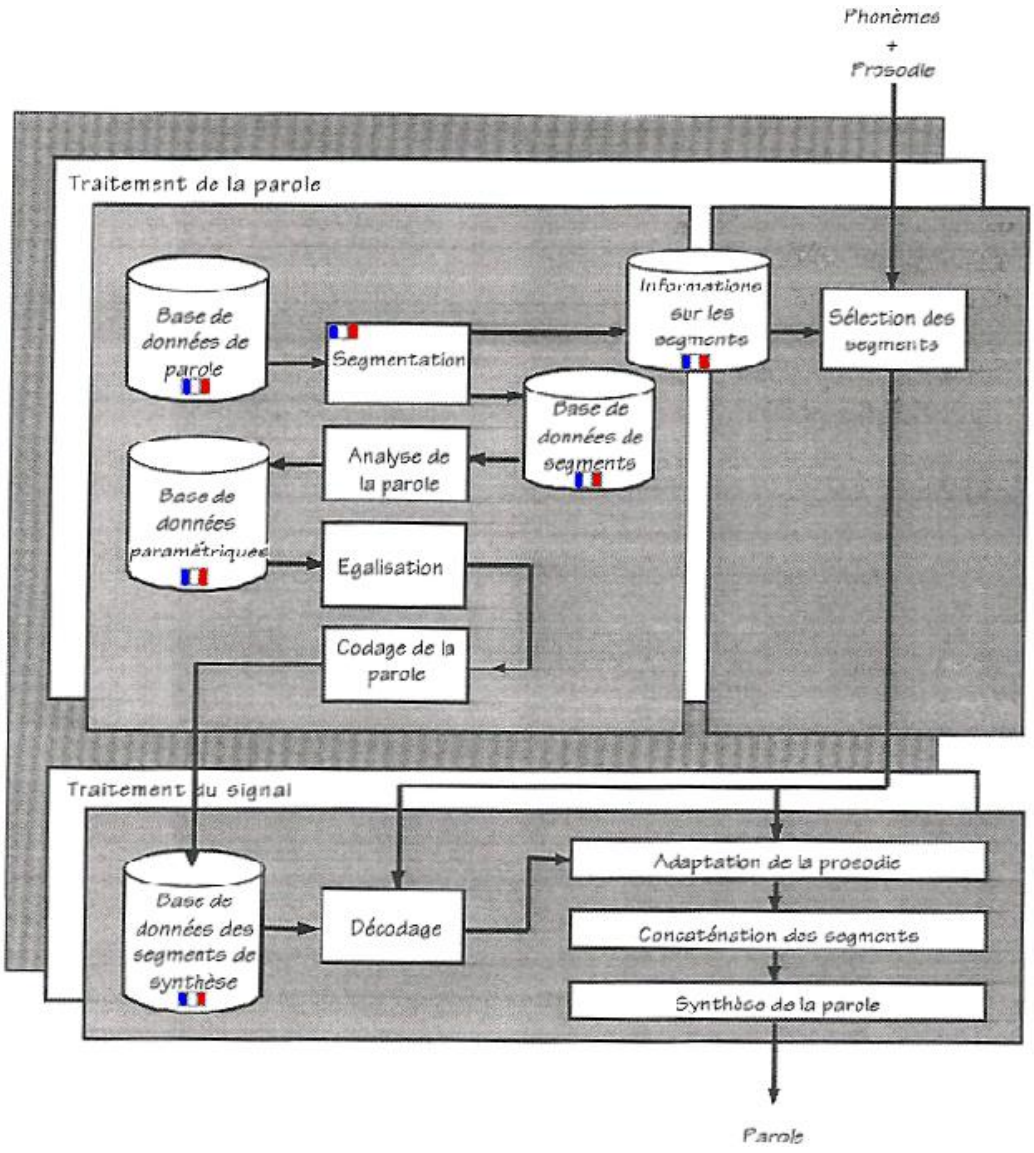


FIGURE III.4 – Schéma général d'un synthétiseur par concaténation. Les opérations qui dépendent de la langue sont indiquées par un drapeau.(d'après [5])

automatiquement grâce à des algorithmes de segmentation automatique dont les décisions sont ensuite vérifiées et éventuellement corrigées manuellement. Le résultat de cette segmentation constitue la base de données de segments, qui comprend les échantillons de tous les segments utilisables. On centralise également l'information relative à ces segments (leur nom, leur durée, leur pitch, et les marqueurs de frontières de phonèmes à l'intérieur des segments) dans une base de données séparée, qui sera utilisée par le bloc de sélection d'unités. Dans le cas de diphones, par exemple, on mémorise l'instant de passage d'un phonème à l'autre afin de pouvoir plus tard modifier séparément les durées de chaque demi-phonème.

Modélisation paramétrique et codage

On soumet souvent le signal de ces unités de parole à une modélisation paramétrique, qui a pour effet de transformer le signal (suite d'échantillons) en une séquence de paramètres d'un modèle, recueillie à la sortie d'un analyseur (Par exemple un modèle LPC, ou un modèle Harmoniques + bruit) et stockée dans une base de données paramétrique. Cette opération rappelle à bien des égards l'analyse menée en synthèse par règles, mais son objectif est ici assez différent. Il ne s'agit pas en effet d'assurer une bonne "interprétabilité" des paramètres du modèle par un phonéticien, mais plutôt de bénéficier des avantages suivants :

- Un modèle bien choisi permet souvent une réduction de la taille des données. On pourra donc se permettre de stocker plus d'unités pour une même quantité de mémoire, ou réduire la taille mémoire nécessaire pour un nombre donné d'unités. Ceci justifie la présence d'un codeur de parole sur la figure III.4. Cet avantage est important en synthèse par concaténation étant donné le grand nombre d'unités à stocker.
- De nombreux modèles de parole séparent explicitement les contributions respectives de la source et du conduit vocal . Ceci est mis à profit par le synthétiseur pour résoudre indépendamment (et donc plus simplement)les deux problèmes fondamentaux évoqués plus haut : la modification de la prosodie des unités et leur concaténation.
- De même, certains modèles séparent explicitement la parole en deux contributions (une contribution voisée et une contribution non-voisée) ce qui permet aussi d'améliorer les problèmes de raccordement des unités au moment de la synthèse

Sélection des segments

Lorsque la base de données des segments de synthèse a été constituée, la synthèse proprement dite peut commencer. Les informations phonétiques et prosodiques présentées à l'entrée du synthétiseur sont tout d'abord transformées en séquences de commandes de segments du synthétiseur. Ceci est réalisé à l'aide du module de sélection de segments (ou d'unités) de la figure III.4. La distinction introduite plus haute entre sélection segmentale d'unités et sélection totale (segmentale et suprasegmentale) est bien entendu d'application ici. En sélection segmentale, la suite des unités à concaténer est déduite de la chaîne phonétique d'entrée uniquement. Au contraire, la sélection totale implique un choix d'unités réalisant au mieux les caractéristiques segmentales et suprasegmentales (typiquement : pitch et durée) de la chaîne phonétique d'entrée. Que ce soit en sélection segmentale ou totale (la première n'étant qu'un cas particulier de la seconde), deux cas de figure peuvent se présenter.

Sélection statique Dans le premier, chacune des unités à synthétiser peut être déduite indépendamment des autres, directement à partir de la suite des phonèmes à produire. C'est le cas par exemple d'une synthèse avec sélection segmentale de diphones dans une base de données ne

contenant qu'une seule instance de chaque diphone : la détermination de chaque diphone ne dépend que d'un couple de phonèmes successifs dans la chaîne phonétique d'entrée. On parle alors de sélection statique. Notons que la plupart des systèmes actuels suivent encore cette stratégie.

Sélection dynamique On considère au contraire la sélection dynamique lorsque le choix de la suite d'unités à concaténer ne peut se faire que par minimisation d'un coût de sélection global sur toute la phrase à synthétiser (auquel cas le choix d'une unité interfère avec le choix d'une autre). C'est le cas des algorithmes de sélection automatique d'unités dites non-uniformes apparus récemment, qui procèdent par sélection totale et dynamique. Au moment de choisir les segments à mettre en oeuvre, plusieurs instances d'une même unité phonétique sont disponibles, avec des prosodies différentes et positionnées (dans le corpus) dans des contextes phonétiques différents. Il faut donc, pour réaliser au mieux la synthèse, choisir les segments dont le contexte est le plus proche de la chaîne phonétique à synthétiser, dont la prosodie se rapproche également le plus de la prosodie à produire, et dont les extrémités ne présentent pas trop de discontinuités spectrales l'une par rapport à l'autre [2]. On procède donc en général par programmation dynamique (algorithme de Viterbi) dans le treillis des segments utilisables, de façon à minimiser :

- le coût de sélection global évoqué plus haut, qui tient compte : du coût de représentation (dans quelle mesure les segments choisis correspondent-ils au contexte phonétique et prosodique dans lequel on les insère ?)
- et le coût de concaténation (dans quelle mesure la juxtaposition des segments choisis amène-t-elle des discontinuités).

Concaténation des segments

Une fois les unités choisies, et après en avoir déduit la prosodie à partir des spécifications prosodiques d'entrée (qui se trouvent être associées à la chaîne phonétique d'entrée), le synthétiseur puise dans la base de données paramétrique pour y extraire les flux paramétriques des unités à juxtaposer. Après les avoir judicieusement décodées, il les envoie à un module de modification de la prosodie qui ajuste le pitch et la durée de chaque unité aux spécifications produites par le module de sélection.

Si les segments sont représentés sous forme paramétrique, cette opération implique typiquement une modification des paramètres associés à la source (d'où l'intérêt des modèles où ces paramètres sont indépendants des paramètres du conduit).

A la sortie du module d'adaptation de la prosodie, les possibles discontinuités de pitch entre segments successifs se trouvent implicitement éliminées. Il reste cependant d'éventuelles discontinuités spectrales. Le rôle du module de concaténation est de les éliminer dans la mesure du possible, par lissage spectral dans le domaine paramétrique. Ici aussi, le choix du modèle utilisé se révèle être de première importance : bien choisi, il permet, par simple lissage temporel linéaire de ses coefficients, de réaliser un lissage spectral qui correspond approximativement au passage naturel d'un son à l'autre (lequel est soumis par nature à des contraintes physiologiques, qu'il n'est pas toujours évident de respecter).

Modification de la fréquence fondamentale et de la durée

La prosodie est réalisée en utilisant des méthodes de modification de la fréquence fondamentale et de la durée des segments. On s'intéresse ici aux méthodes permettant de réaliser

2. notons ici que si la phrase à synthétiser est entièrement contenue dans le corpus, l'unité choisie peut être cette phrase elle même annulant de fait tout problème de concaténation

indépendamment une modification de l'échelle temporelle ou fréquentielle d'un signal :

- La modification de l'échelle temporelle permet d'altérer arbitrairement la durée d'un signal sans en modifier (si possible) le contenu fréquentiel.
- La modification de l'échelle fréquentielle est l'opération duale de la précédente, et consiste à modifier la hauteur d'un son donné, sans en modifier la durée. En traitement de la parole on désire obtenir un changement de hauteur tonale tout en conservant la position des formants.

Les méthodes permettant de réaliser une modification de l'échelle temporelle ou fréquentielle se répartissent en deux catégories :

- Les méthodes paramétriques, reposant sur un modèle de signal précis (par exemple, modèle sinusoïdal),
- Les méthodes non-paramétriques (où il n'est fait aucune hypothèse sur la nature du signal traité). Les méthodes non-paramétriques peuvent se répartir à nouveau en deux catégories : les méthodes travaillant dans le domaine temporel et les méthodes travaillant dans le domaine fréquentiel.

dans ce cours nous ne détaillerons qu'une méthode non-paramétrique (très utilisée) travaillant dans le domaine temporel : la méthode TD-PSOLA (pour "*Time Domain Pitch synchronous OverLap and Add*")

Méthode temporelle TD-PSOLA [3] Cette méthode suppose que l'on traite un signal de parole dont on connaît la période fondamentale. L'idée [20] est encore fondée sur l'hypothèse que le signal de parole est constitué d'impulsions glottales filtrées par le conduit vocal. On observe ainsi une succession de réponses impulsionnelles, positionnées à des temps multiples de la période (hypothèse du peigne temporel convolué avec la réponse impulsionnelle du conduit vocal).
On définit d'abord des 'marques d'analyses' synchrones de la fréquence fondamentale pour les parties voisées, positionnées sur la forme d'onde à chaque période. Les modifications d'échelles sont alors effectuées de la façon suivante :

Modification de l'échelle temporelle Pour modifier la durée du signal sans en altérer la fréquence fondamentale, on va simplement dupliquer (étirement temporel) ou éliminer (compression temporelle) des périodes de la forme d'onde, en fonction du taux de modification désiré. On est donc conduit à définir des marques de synthèse également synchrones du fondamental, associées aux marques d'analyse (de façon non-bijective puisque certaines marques sont dupliquées ou éliminées).

Les signaux à court-terme situés autour de chaque marque d'analyse sont alors extraits (par l'utilisation d'une fenêtre temporelle-par exemple de type hanning- de durée égale à deux périodes et centrée sur la marque d'analyse) et 'recopiés' autour des marques de synthèse correspondantes et le signal modifié est obtenu par une simple méthode d"overlap/add". La figure III.5 illustre le principe de cette méthode pour un taux d'étirement temporel local de 1.5.

On voit que deux périodes du signal original ont donné naissance à trois périodes dans le signal modifié, ce qui correspond bien à un étirement temporel mais la durée de la période n'est pas modifiée (l'écartement des marques de synthèse est le même que celui des marques d'analyse), la fréquence fondamentale du signal est conservée.

La figure III.6 donne un exemple d'application à la phrase 'il s'est' dont l'original est donné en haut de la figure. On remarque la partie non-voisée au centre de la fenêtre (le son 's'), séparant les deux parties voisées /i/ et /e/.

3. Paragraphe écrit par J. Laroche [18]

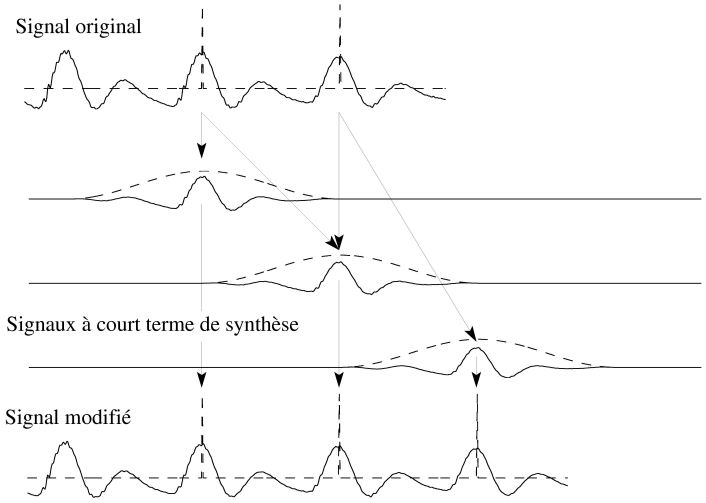


FIGURE III.5 – Modification de la durée du signal par la méthode TD-PSOLA. En haut, le signal original, au milieu trois signaux à court-terme générés à partir des deux signaux à court-terme centrés autour des deux premières marques d'analyse. En bas, signal modifié.

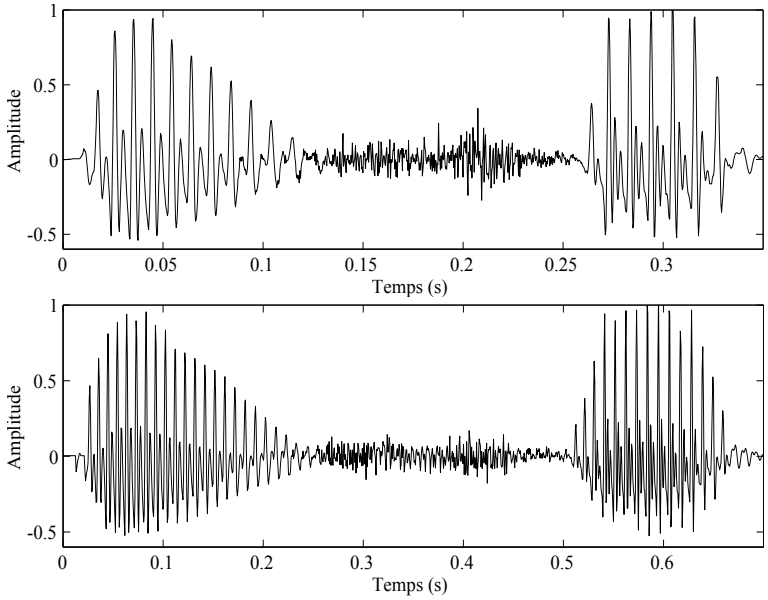


FIGURE III.6 – Original : "il s'est"(d'après [18])

Modification de l'échelle fréquentielle Si l'on est capable de positionner dans le signal les marques d'analyse exactement sur le début de chaque onde glottale (réponse impulsionnelle du conduit vocal se produisant à chaque fermeture glottale), on conçoit que diminuer (resp. augmenter) l'intervalle de temps séparant deux marques d'analyse consécutives va permettre d'augmenter (resp. de diminuer) la fréquence du fondamental, sans que les formants soient modifiés (la réponse impulsionnelle n'est pas modifiée, en particulier sa décroissance temporelle et ses fréquences de résonance–les formants).

On est ainsi conduit à définir des marques de synthèse correspondant à la valeur modifiée du fondamental, et à les associer aux marques d'analyse comme précédemment. Puisque les marques de synthèse sont plus serrées (élévation du fondamental) ou écartées (abaissement du fondamental) que dans le signal original, il faut pour conserver la durée du signal dupliquer ou éliminer certaines marques. La figure III.7 illustre le principe de cette méthode.

On constate que les marques de synthèse étant plus écartées que les marques d'analyse, la période du signal est allongée. Pour éviter une élongation du signal, il est nécessaire d'éliminer périodiquement certains signaux à court-terme.

Lorsque le signal ne possède plus de fréquence fondamentale bien précise (cas des consonnes etc...), la modification est réalisée de façon non-synchrone, jusqu'à ce que l'on retrouve une région présentant un fondamental plus net.

La méthode décrite ci-dessus réalise des modifications de très bonne qualité. Par sa simplicité, elle peut faire l'objet d'une implémentation temps réel. Par contre, il est important de noter que la qualité des modifications de fondamental sont très sensibles à la position des marques d'analyse. On peut alors se tourner vers d'autres méthodes (notamment fréquentielles). Pour d'autres méthodes basées sur des idées très similaires, on pourra se référer à [21], ou à l'article de J. Laroche dans [17].

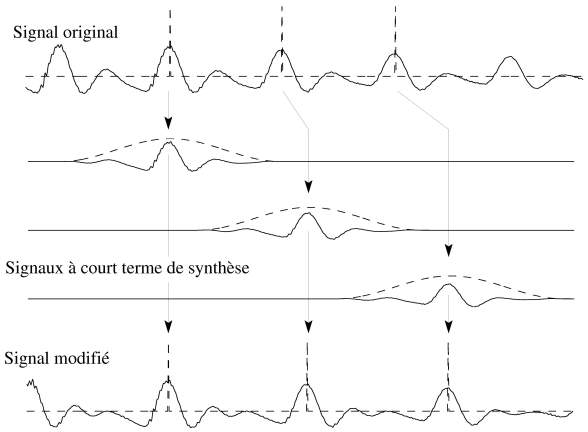


FIGURE III.7 – Modification de la hauteur du signal par la méthode TDPSOLA. En haut, le signal original, au milieu trois signaux à court-terme générés à partir des trois premières marques d'analyse. En bas, signal modifié. L'écartement des marques de synthèse n'est pas identique à celui des marques d'analyse.

Notons enfin, que la méthode PSOLA peut être appliquée sur le signal d'excitation qui aurait été préalablement obtenu à l'aide d'un modèle paramétrique (par exemple par prédiction linéaire, ce qui donnera la méthode LP-PSOLA).

Bibliographie

[1] J. Allen, S. Hunnicut, and D. Klatt. *From Text To Speech, The MITALK System.* Cambridge University Press, Cambridge, 1987.

[2] F. Beaugendre. Modèles de l'intonation pour la synthèse. http ://www.bibliotheque.refer.org/parole/beaugend/beaugend.htm#F2, 1995.

[3] B. Bleicher. Anatomie de l'appareil respiratoire et des mécanismes phonatoires. http ://gillesdenizot.com/fr/articles/Anat-physio.pdf, 2001.

[4] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and H. Leich. *Traitement de la parole.* Presses polytechniques et universitaires romandes,Lausanne, 2000.

[5] Calliope. *La parole et son traitement automatique.* Collection CNET - ENST. Masson, 1989.

[6] Michelle Crank, Brain-Goddess, Debby Lee, Sophie Kallinis, and Adam Friedman. Anatomy. WWW, 2000. http ://www.molbio.princeton.edu/courses/mb427/2000/projects/0008/anatobrain.html.

[7] Le Monde de l'APNÉE. L'apareil respiratoire anatomie et généralités. http ://www.chez.com/defaut/apnee/anatresp.html, 1997.

[8] T. Dutoit. Introduction au traitement automatique de la parole, notes de cours ; dec2. http ://tcts.fpms.ac.be/cours/1005-08/speech/, 2000.

[9] G. Fant. *Acoustic theory of Speech Production.* Mouton, La Hague, 1960.

[10] J. Flanagan. *Speech Analysis, Synthesis and Perception.* Spinger Verlag,Berlin, 1972.

[11] IPA. International phonetic alphabet. http ://www2.arts.gla.ac.uk/IPA/ipachart.html.

[12] M. Kahrs and K. Brandenbourg. *Applications of Digital Signal Processing to Audio and Acoustics.* Kluwer Academic Press, Dortrecht, Netherland,, 1998.

[13] J. Laroche. Traitement des signaux audiofréquences. Technical report, E.N.S.T. Polycopié de cours.

[14] Theodore Levin and Michael Edgerton. Le chant des touvas. *Pour la science*, (N° 265), Novembre 1999. http ://www.pourlascience.com/numeros/pls-265/art-5.htm.

[15] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6) :453–467, Dec 1990.

[16] E. Moulines and J. Laroche. Non parametric techniques for pitch-scale and time-scale modification of speech. 16 :175–205, Feb 1995.

[17] Randall L. Plant. Eastern virginia medical school : A web site devoted to describing disorders of the voice and the larynx. http ://www.voice-center.com/index.html, 2001.

[18] L. Rabiner and B. Juang. *Fundamentals of Speech recognition.* Signal prcessing series. Prentice Hall, a. oppenheim,series editor edition, 1993.

[19] K. Stevens. Airflow and turbulence noise for fricative and stop consonants : Static considerations. *J. of Acoust. Soc. Amer.*, 50(2) :1180–1192, 1971.