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

Gaël RICHARD

30 novembre 2012

Table des matières

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

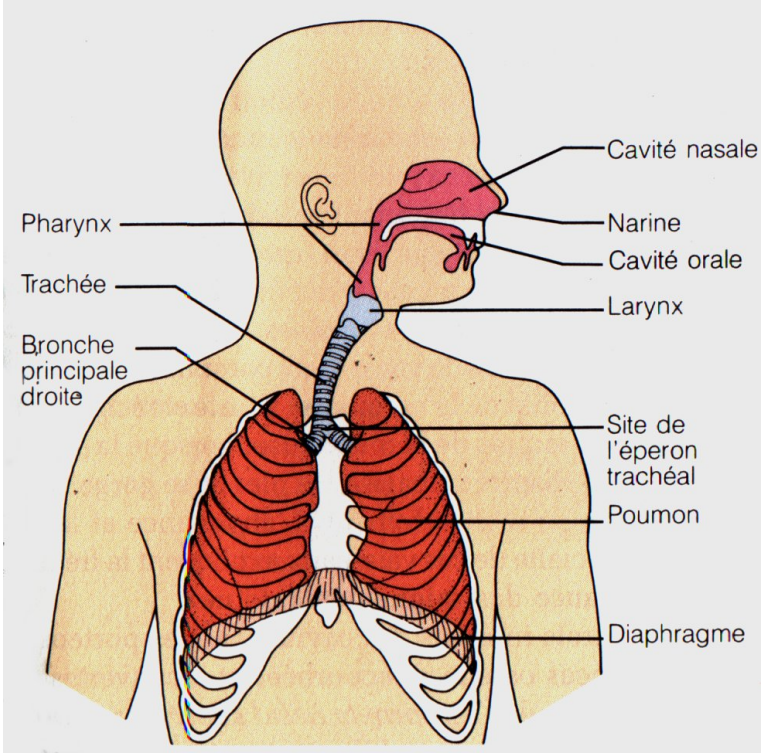


FIGURE I.1 – Schéma de l'appareil respiratoire (d'après [9])

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].

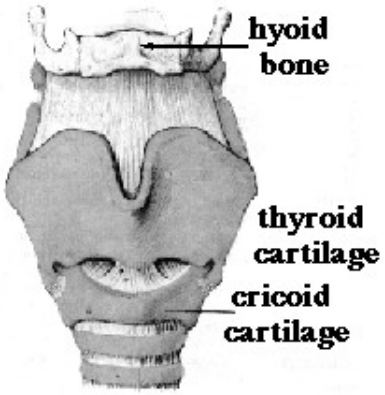


FIGURE I.2 – Schéma du larynx (d'après [21])

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

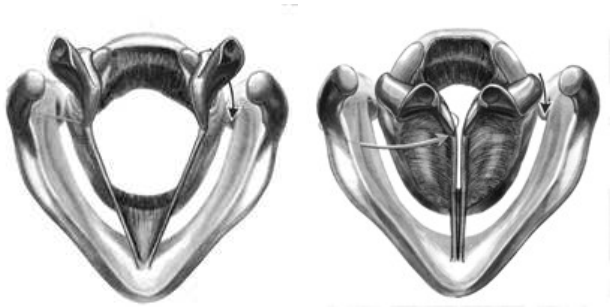


FIGURE I.3 – Les cordes vocales en position ouvertes durant la respiration (à gauche) et fermées pour la production de parole (à droite),(d'après [21])

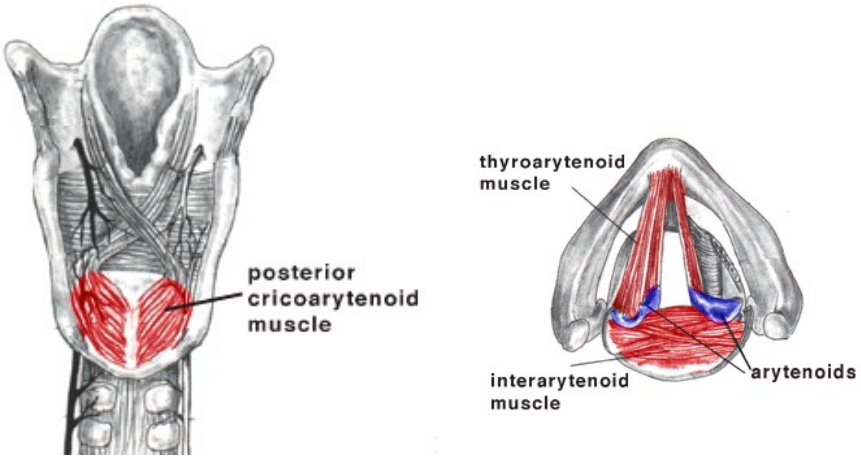


FIGURE I.4 – Schéma des muscles intrinsèques du larynx (d'après [21])

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.

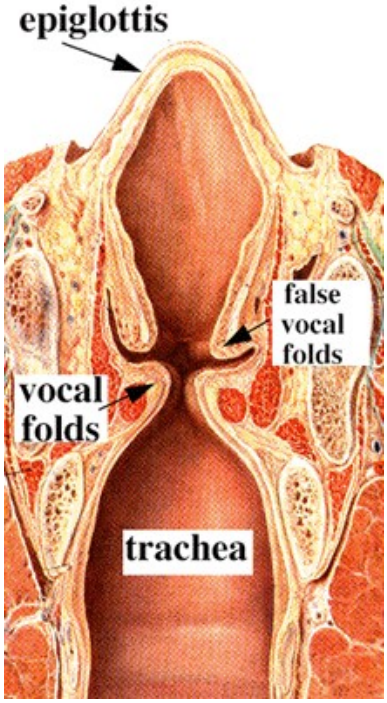


FIGURE I.5 – Vue longitudinale du larynx (d'après [21])

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

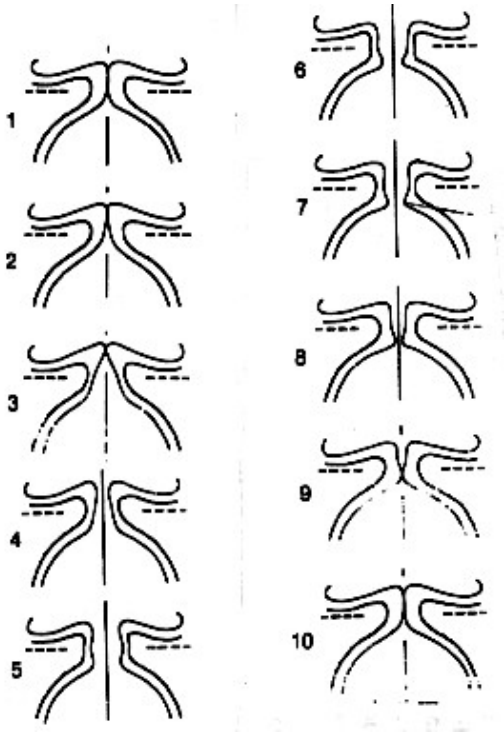


FIGURE I.6 – Schéma de vibration des cordes vocales (d'après [21])

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 [12] 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 ([23], [11]).

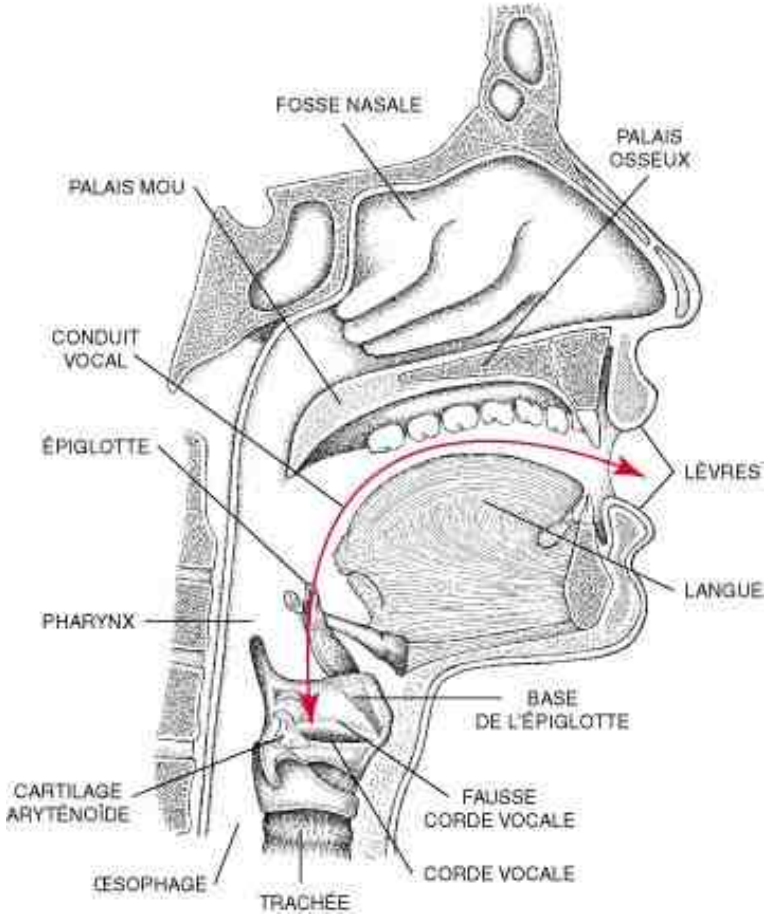


FIGURE I.7 – Vue schématique du conduit vocal humain d'après [18]

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 [15]) 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.

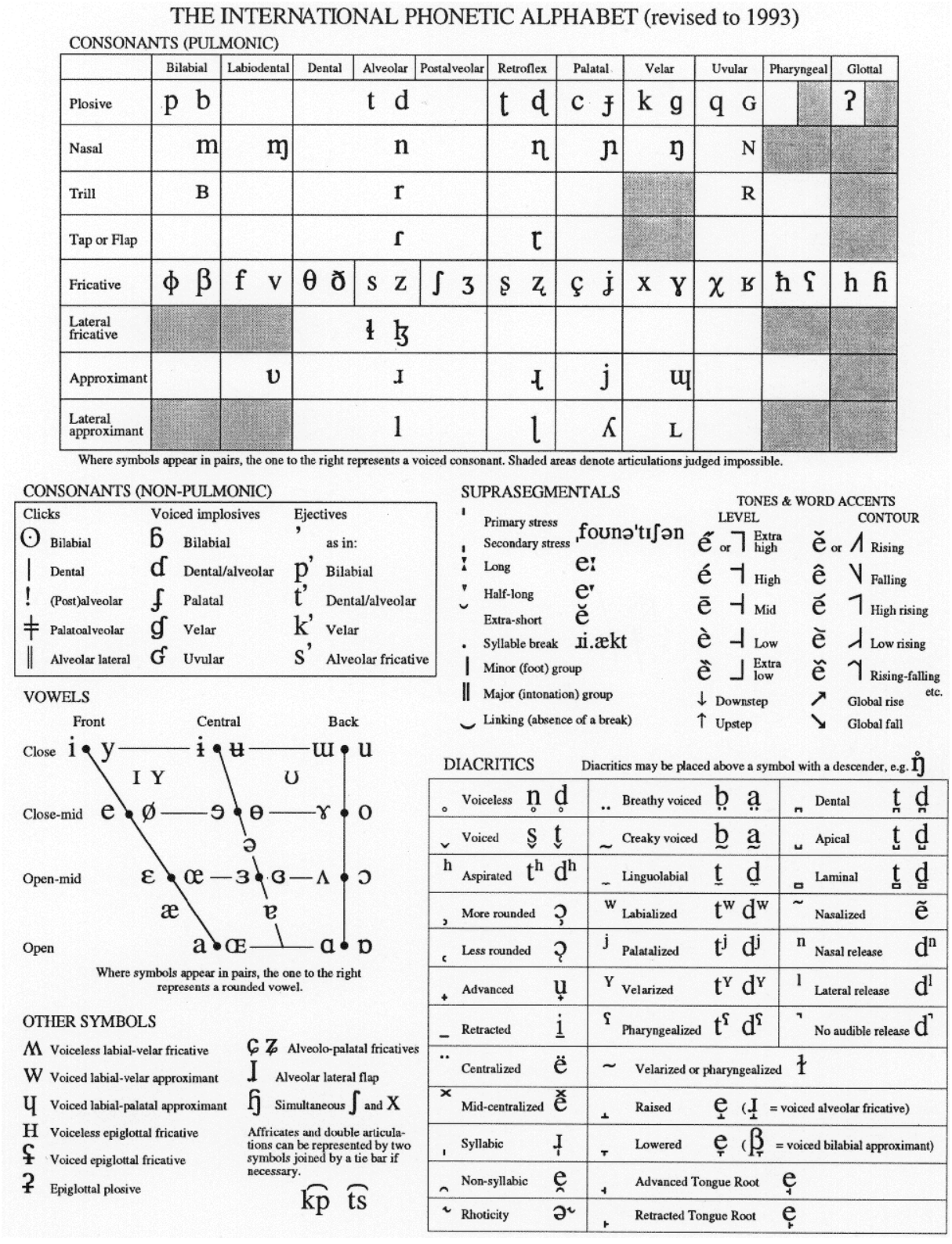


FIGURE I.8 – Tableau complet de l'alphabet phonétique international [15]

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 :

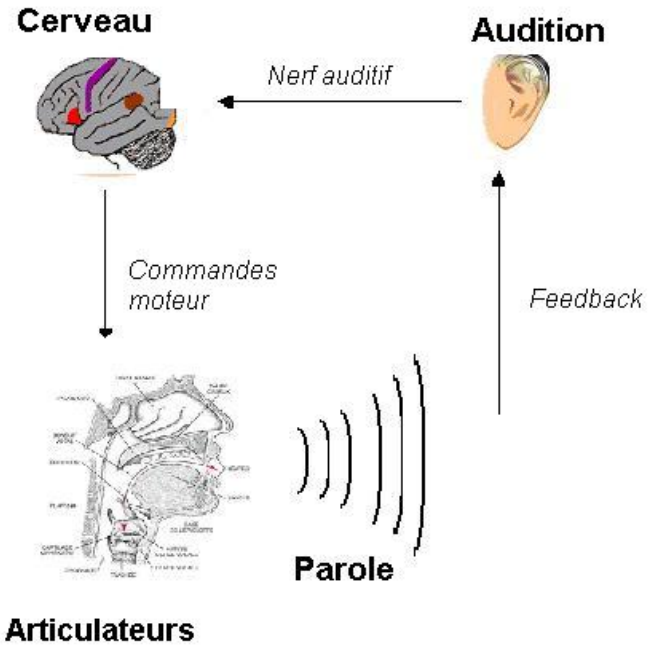


FIGURE I.10 – Système de production et feedback auditif

– 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). ([8])

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

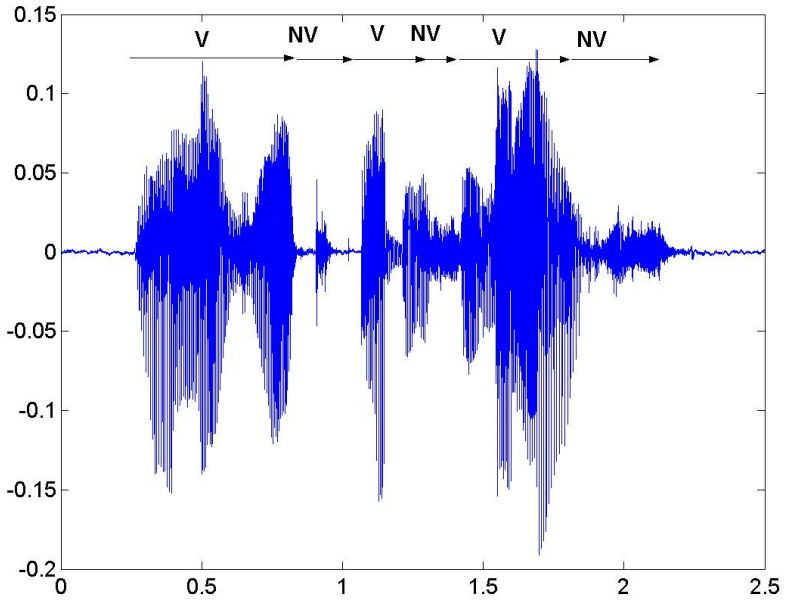


FIGURE I.11 – Signal temporel de la phrase "La musique adoucit les moeurs" : (V=partie voisée ; NV = partie non voisée

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

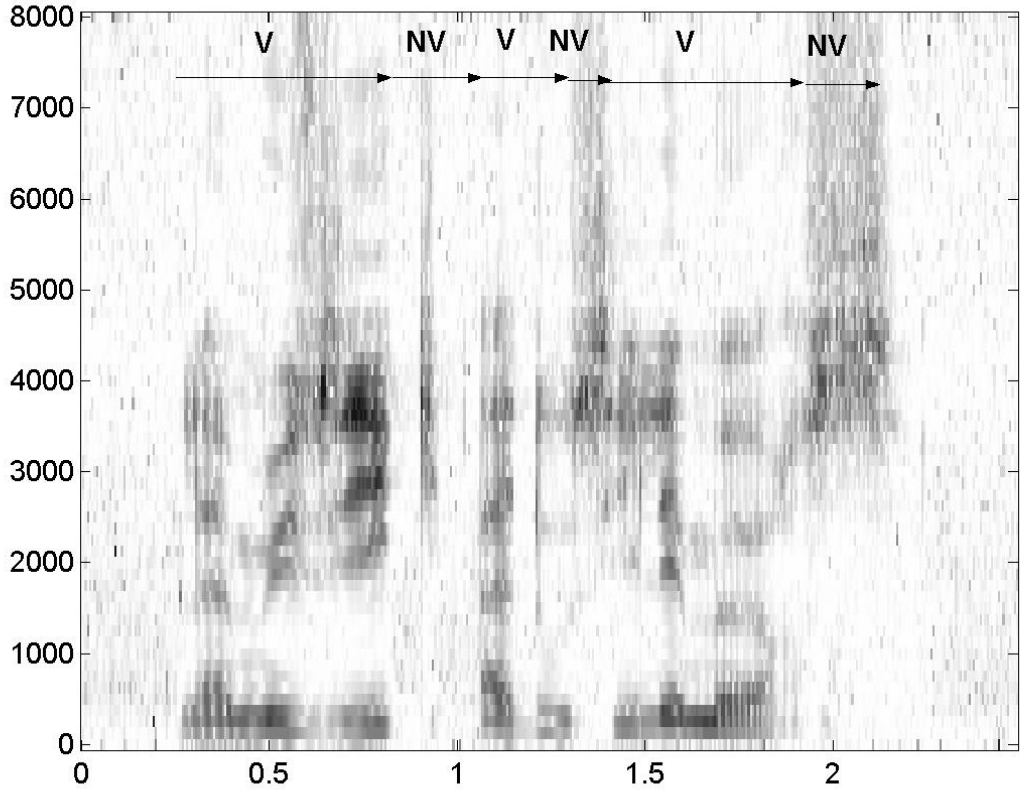


FIGURE I.12 – Spectrogramme de la phrase "La musique adoucit les moeurs" (Le spectrogramme représente le module de la transformée de Fourier où cours du temps avec les Fréquences en ordonnée, le temps en abscisse et l'énergie en niveau de gris. Ainsi une zone sombre, indique une forte énergie à la fréquence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

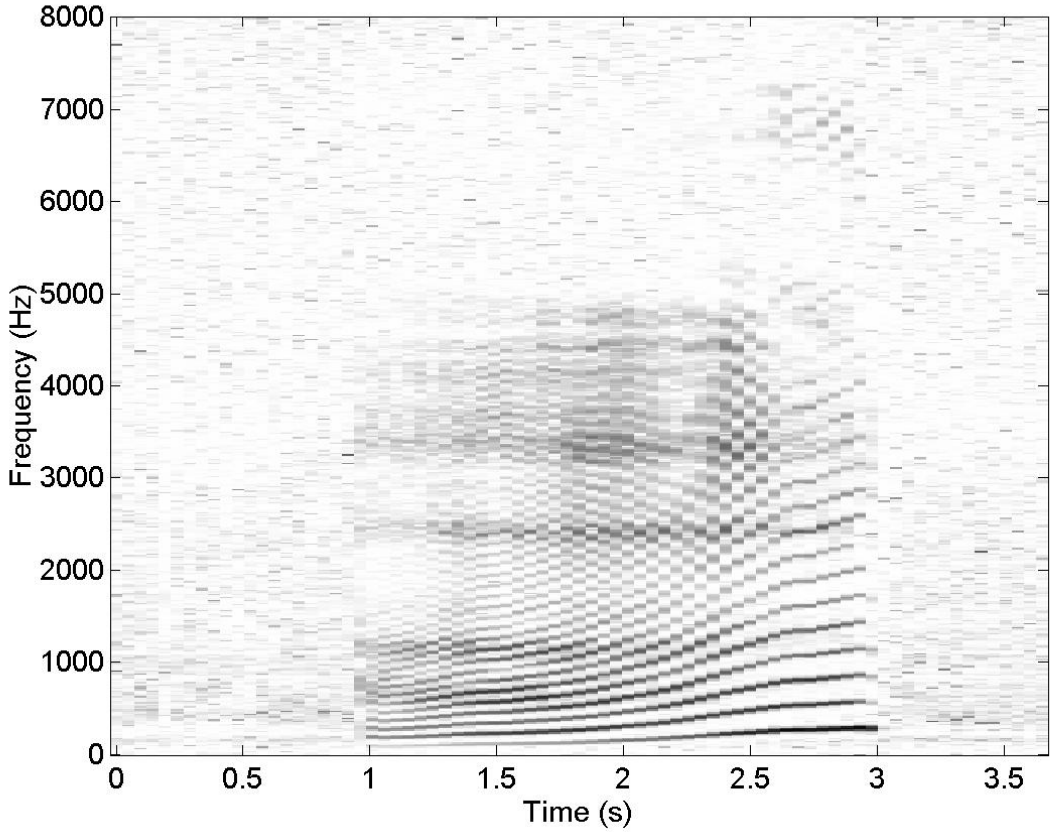
Sur ce spectrogramme, les mouvements brusques de ces formants, notamment les deux premiers, indiquent un changement de voyelle. Comme on le verra plus tard, on peut en effet caractériser les voyelles par la position de leurs seuls deux premiers formants. On ne tient pas compte en général du pic de très 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.



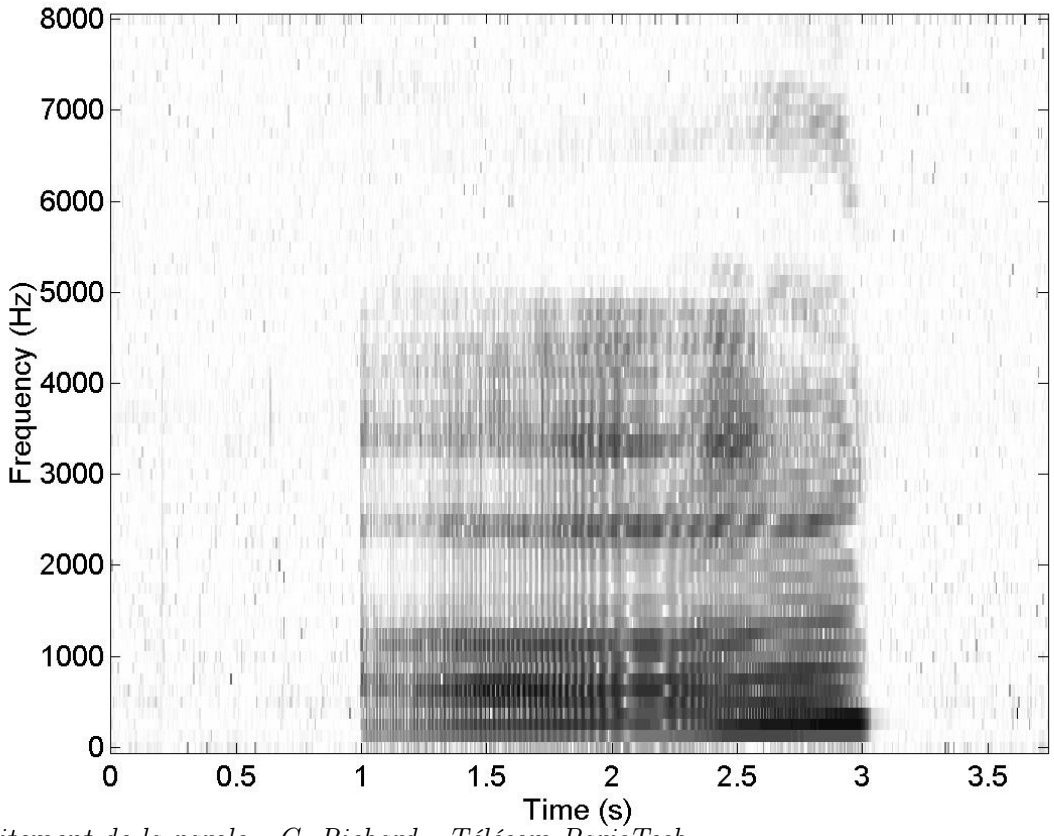


FIGURE I.13 – Spectrogramme bande étroite (haut) et spectrogramme large bande (bas) d'une voyelle /a/ produite avec une élevation progressive de la fréquence fondamentale"

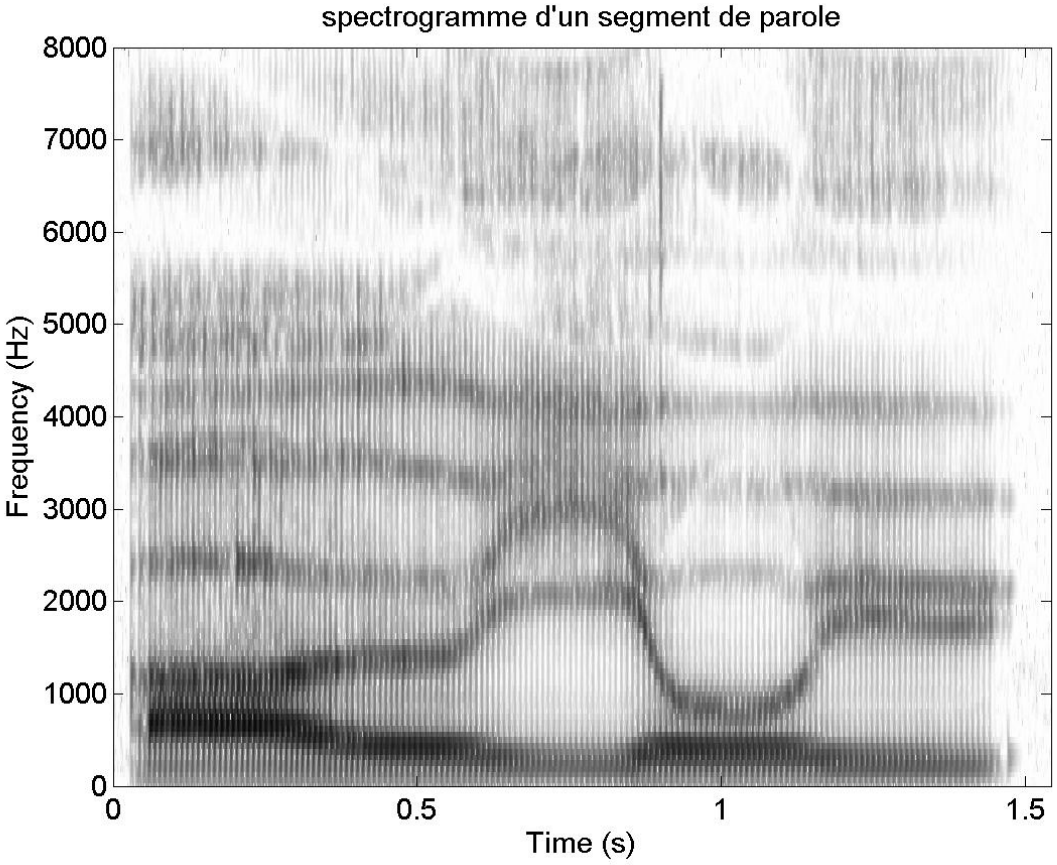


FIGURE I.14 – Spectrogramme du son constitué des voyelles /aeiou/ : les mouvements brusques des deux premiers formants, indiquent un changement de voyelle

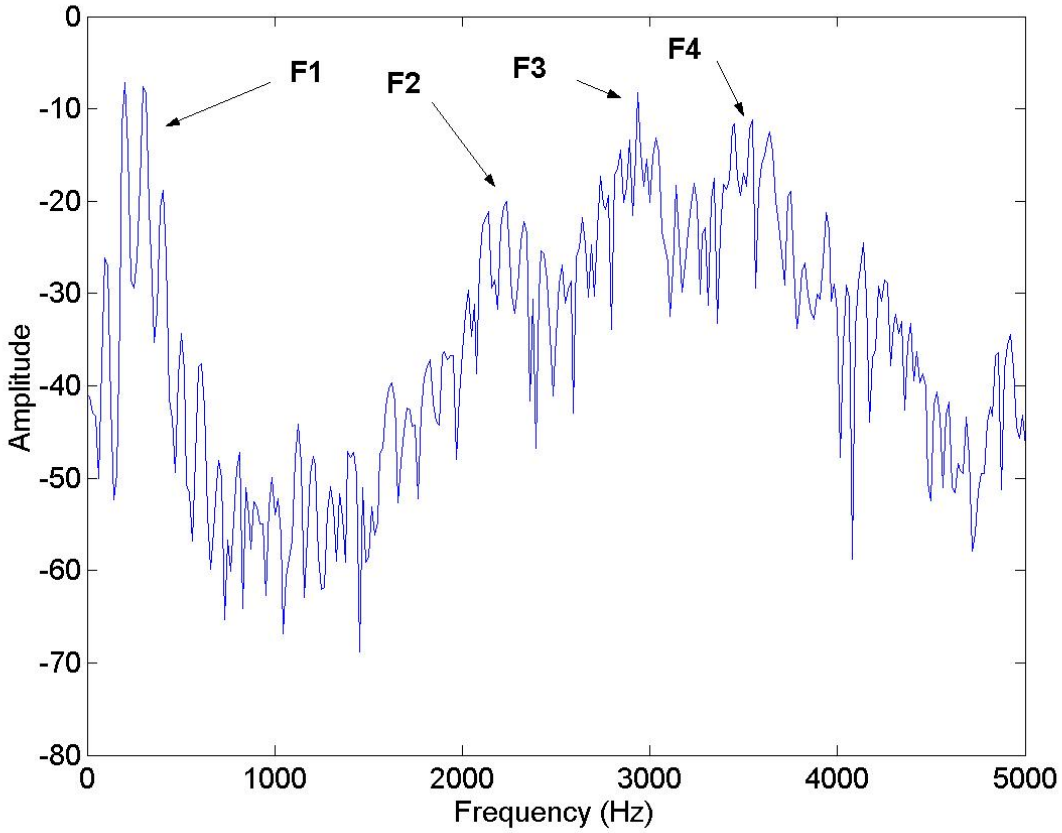


FIGURE I.15 – Module de la transformée de Fourrier (ou coupe spectrographique) d'une trame de la voyelle /i/

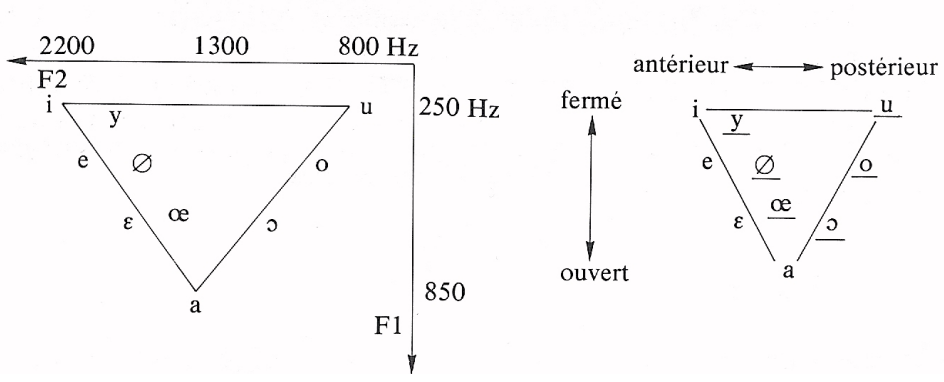


FIGURE I.16 – Triangle vocalique

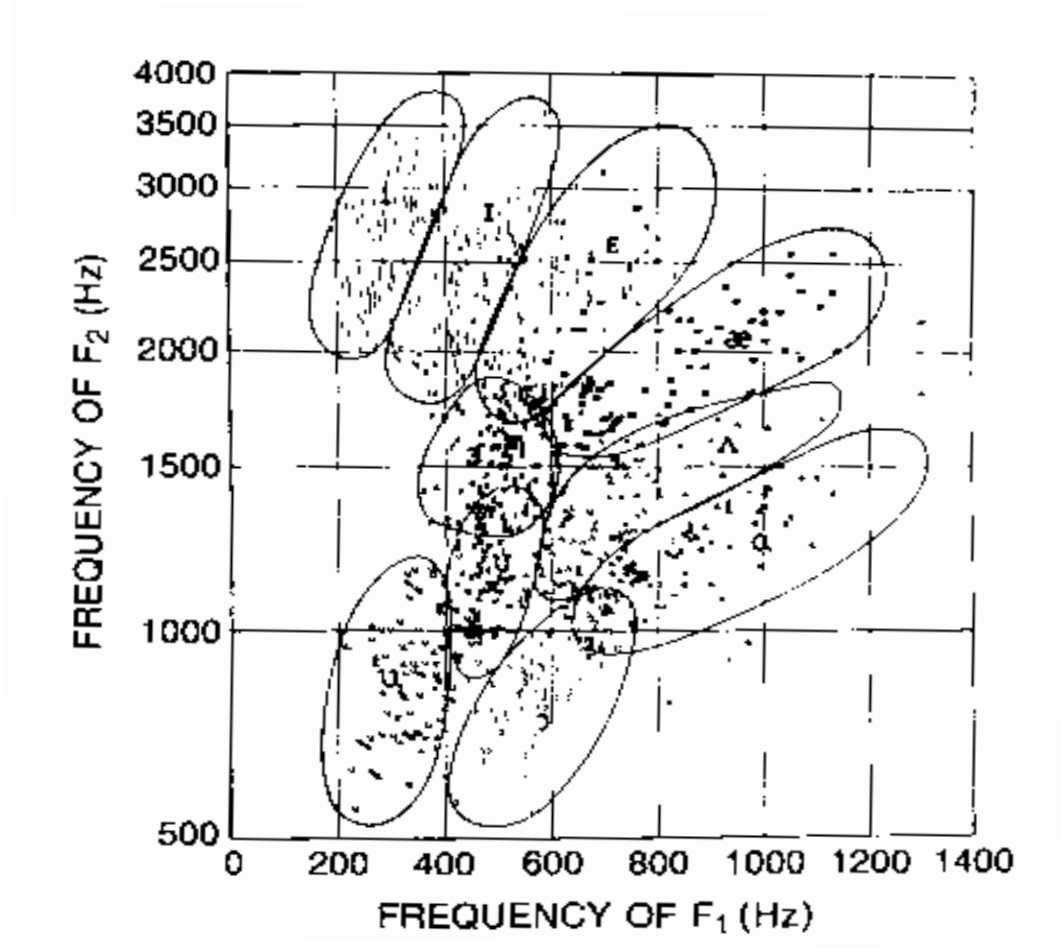


FIGURE I.17 – Représentation des sons vocaliques de l'anglais en fonctions des deux premières fréquences formantiques (D'après [22])

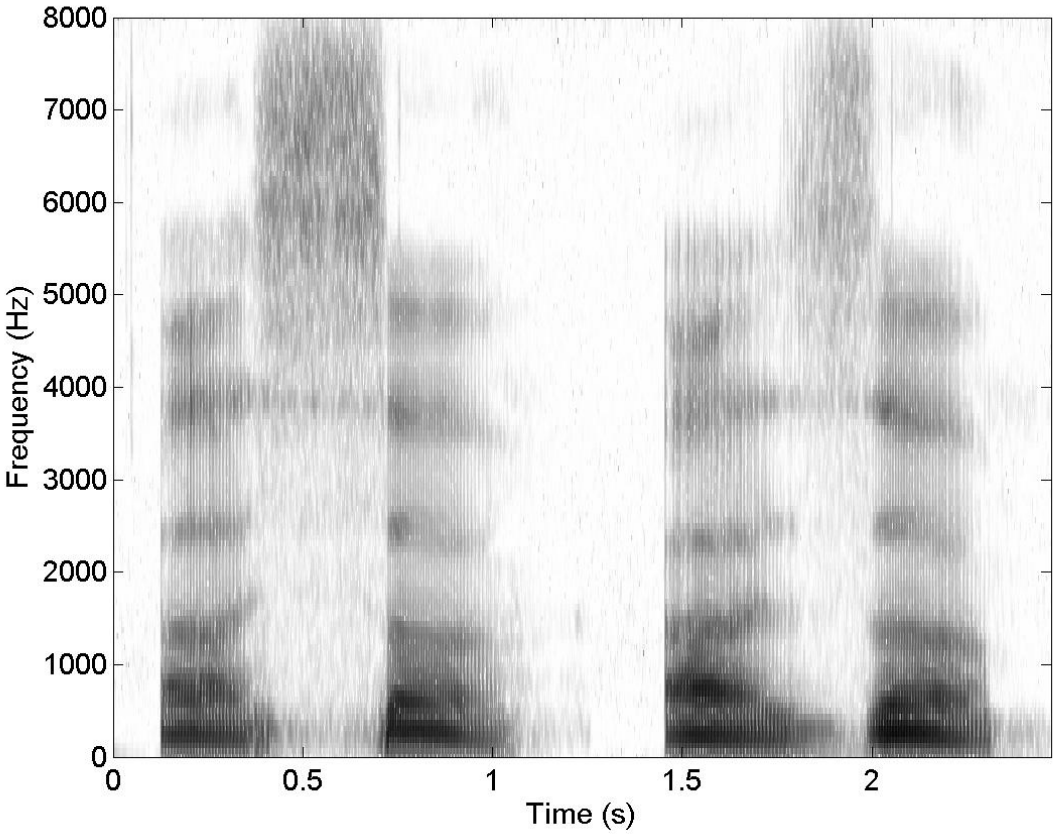


FIGURE I.18 – Spectrogramme bande large du signal "assa aza" (*/a s a a z a/*)"

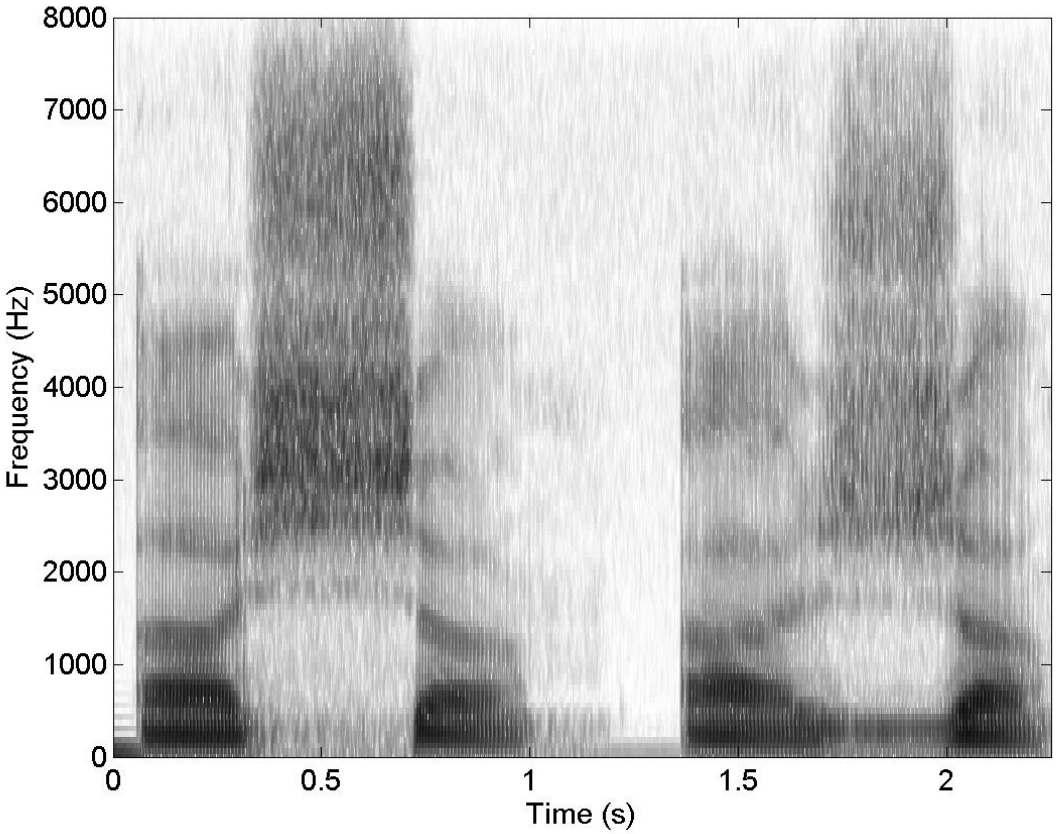


FIGURE I.19 – Spectrogramme bande large du signal "acha aja" (*/a ʃa a ʒa/*)"

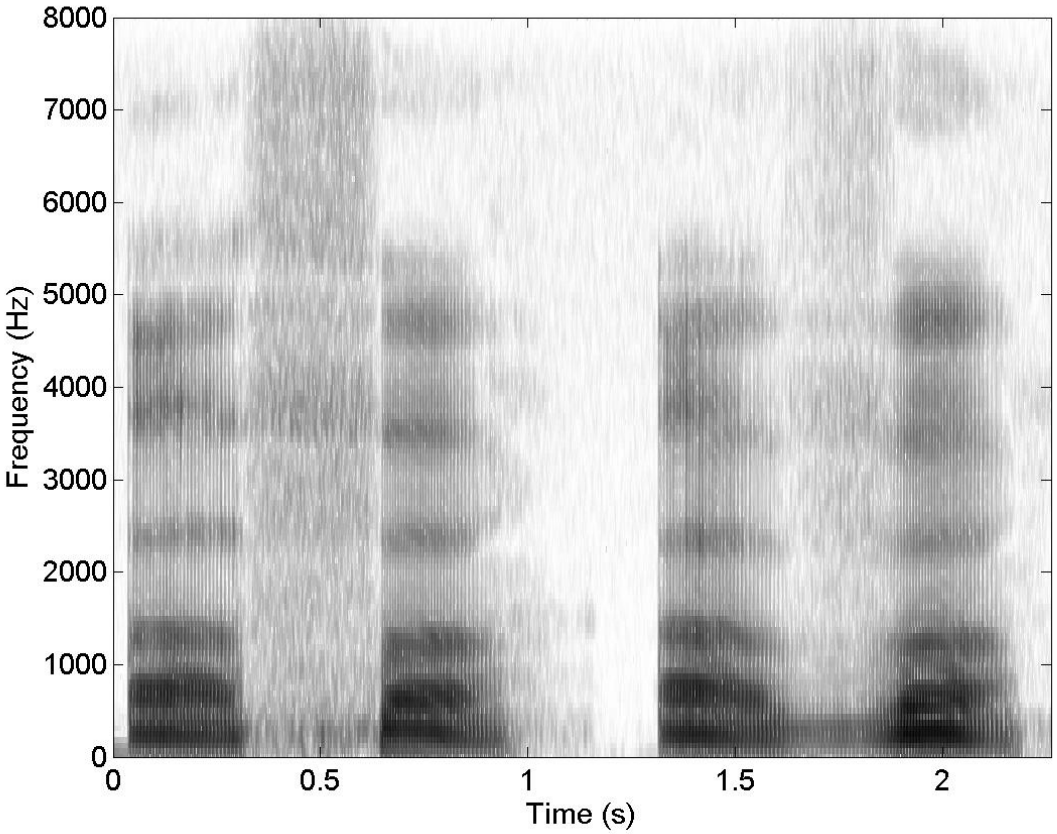


FIGURE I.20 – Spectrogramme bande large du signal "afa ava" (*/a f a a v a/*)"

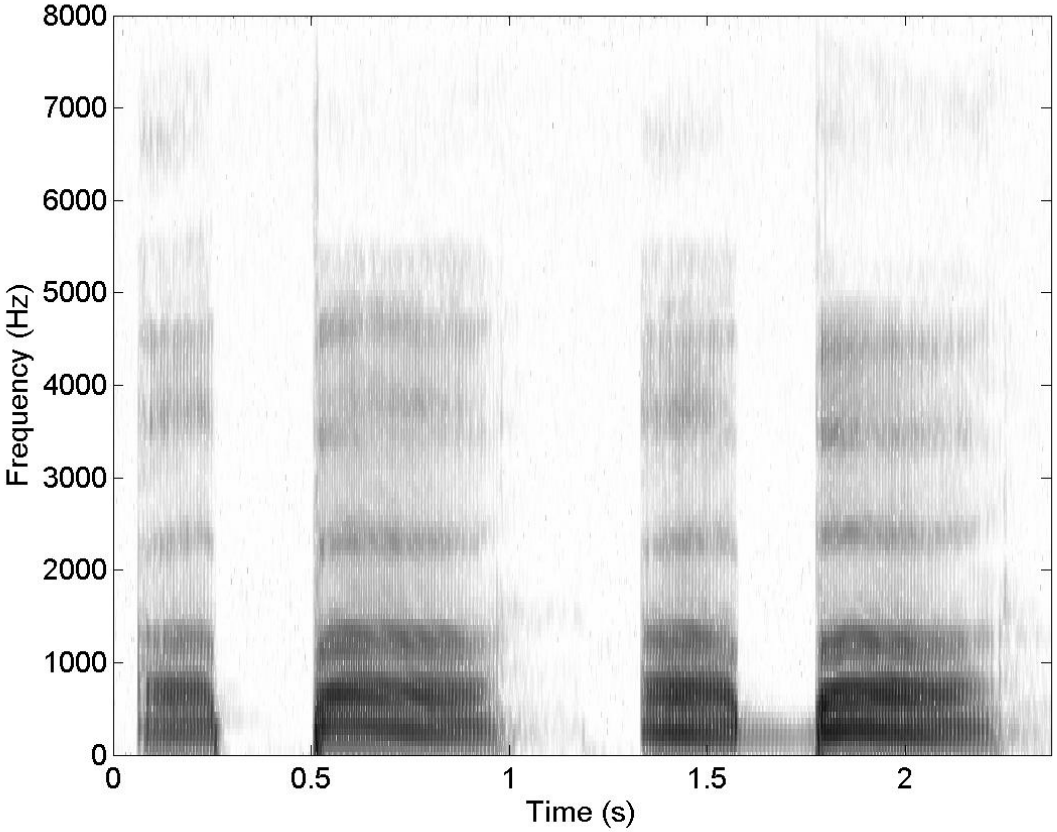


FIGURE I.21 – Spectrogramme bande large du signal "apa aba" (*/a p a a b a/*)"

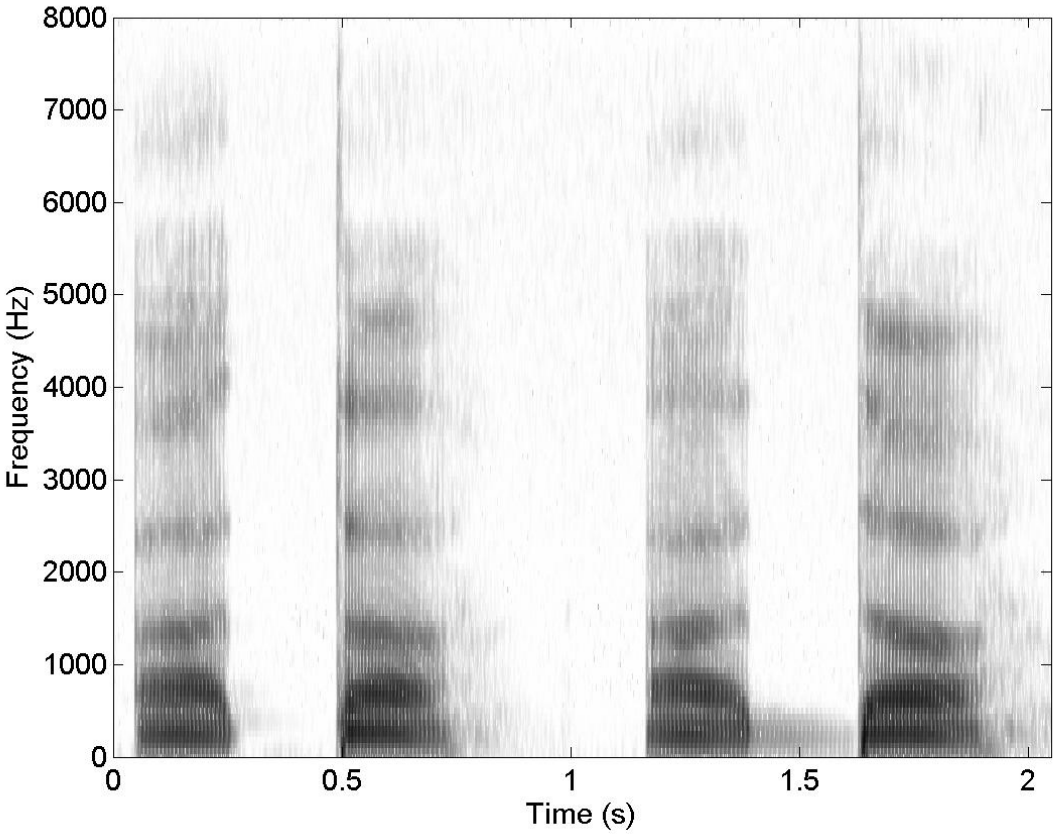


FIGURE I.22 – Spectrogramme bande large du signal "ata ada" (*/a t a a d a/*)"

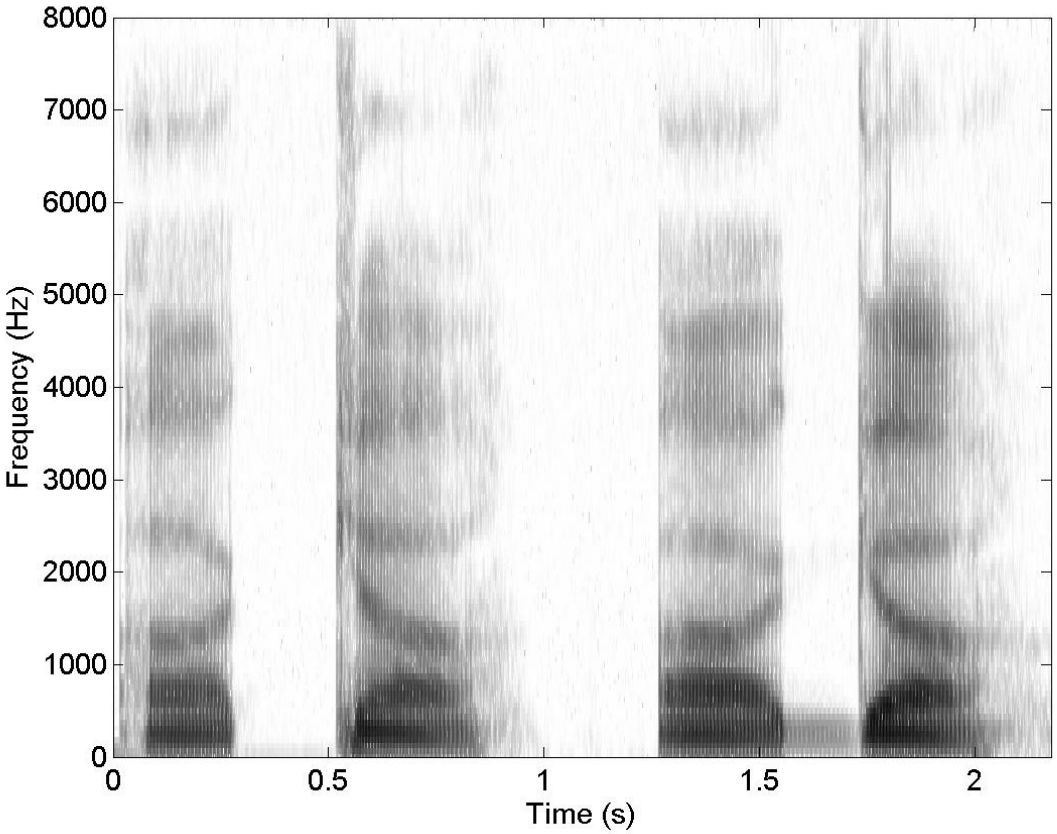


FIGURE I.23 – Spectrogramme bande large du signal "aka aga" (*/a k a a g a/*)"

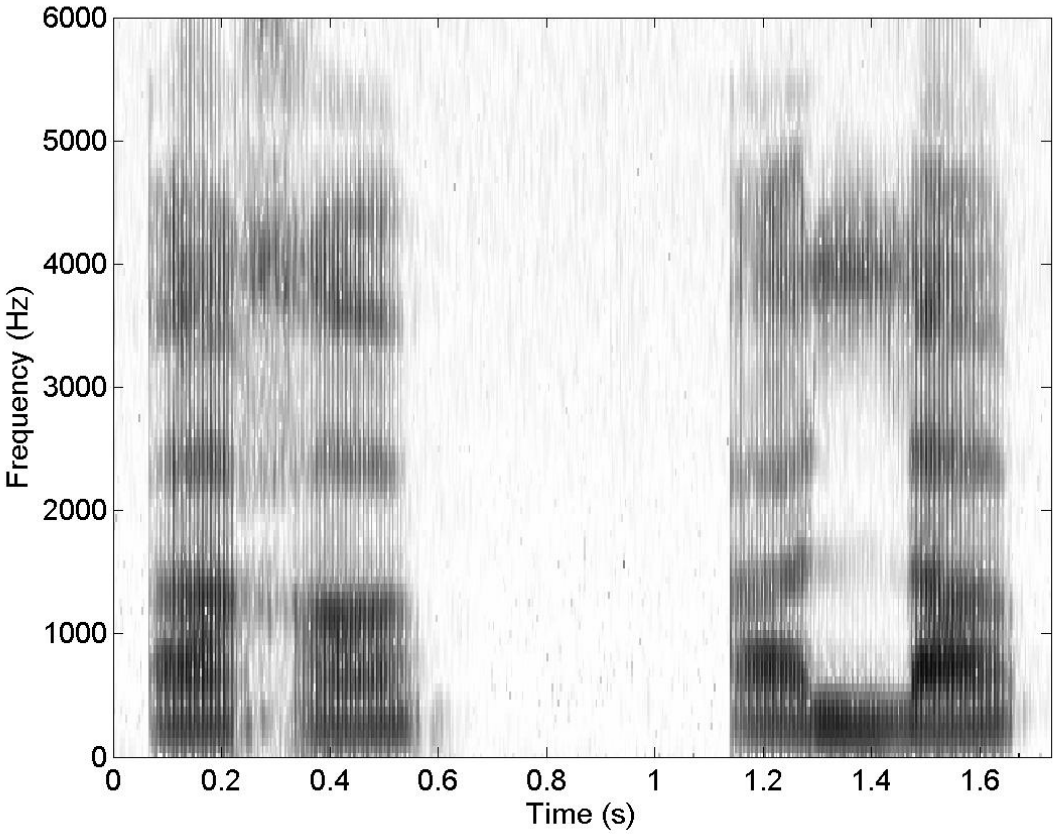


FIGURE I.24 – Spectrogramme bande large du signal "ara ala" (*/a R a a l a/*)"

Chapitre III

Reconnaissance de la parole

III.1 Introduction

La reconnaissance de la parole a pour objectif d'extraire une information lexicale (mots, suite de mots ou hypothèses de mots) à partir d'une information acoustique (le signal de parole).

La compréhension de la parole (pour des applications de dialogue naturel) essaye d'extraire en plus une information sémantique qui permet d'avoir une connaissance des intentions de l'utilisateur. Ces intentions sont formulées sous forme de *concepts*.

La reconnaissance de la parole, étudiée depuis plus de quarante ans, a réalisé des progrès importants et de nombreux systèmes sont maintenant disponibles. Les applications de la reconnaissance vont du "petit" moteur de reconnaissance de quelques mots intégré sur des téléphones portables jusqu'aux applications de dictée vocale avec des vocabulaires de plus de 250 000 mots et aux systèmes de compréhension du langage naturel (pour des applications ciblées).

Malgré les énormes progrès réalisés, il existe toujours un certain nombre d'obstacles pour obtenir des systèmes robustes avec des taux d'erreurs qui seraient comparables à ceux réalisés par l'homme dans la compréhension de la parole naturelle.

Il est clair que ces obstacles prennent leur origine d'une part dans la complexité du signal de parole. Nous avons vu au chapitre II des éléments de production qui permettent de se rendre compte de la complexité de l'appareil phonatoire humain et de la difficulté d'en trouver des modèles. En reconnaissance, il existe un problème supplémentaire qui est lié au fait qu'il n'existe pas un appareil phonatoire humain unique et universel, mais qu'au contraire chaque homme possède des cordes vocales et un conduit vocal uniques qui peuvent s'avérer très différents de ceux de son voisin. Il est ainsi probable que le signal de parole tel qu'il sera capté par un microphone renfermera une grande variabilité suivant les personnes.

On s'aperçoit en fait qu'il existe plusieurs niveaux de variabilité qui peuvent être énumérés ci-dessous :

- *La variabilité intra-locuteur* : qui représente la variabilité de la parole d'un même locuteur au cours du temps. Cette variabilité dépend d'un grand nombre de paramètres tels que la force de la voix, l'état physique (voix enrouée) et de l'état émotionnel (fatigue, colère, excitation, ...). Un exemple de variabilité intralocuteur est donné figure III.1.
- *La variabilité interlocuteur* : qui représente la variabilité entre les différents locuteurs dues aux différences physiologiques, de style d'élocution, d'accents régionaux,etc ... Un exemple de variabilité intralocuteur est donné figure III.2.

Il existe une autre variabilité, pas nécessairement attachée au locuteur, qu'il est particulièrement important en reconnaissance de prendre en compte est celle liée à l'environnement et aux conditions d'enregistrement. Les conditions optimales pour la reconnaissance vocale sera un

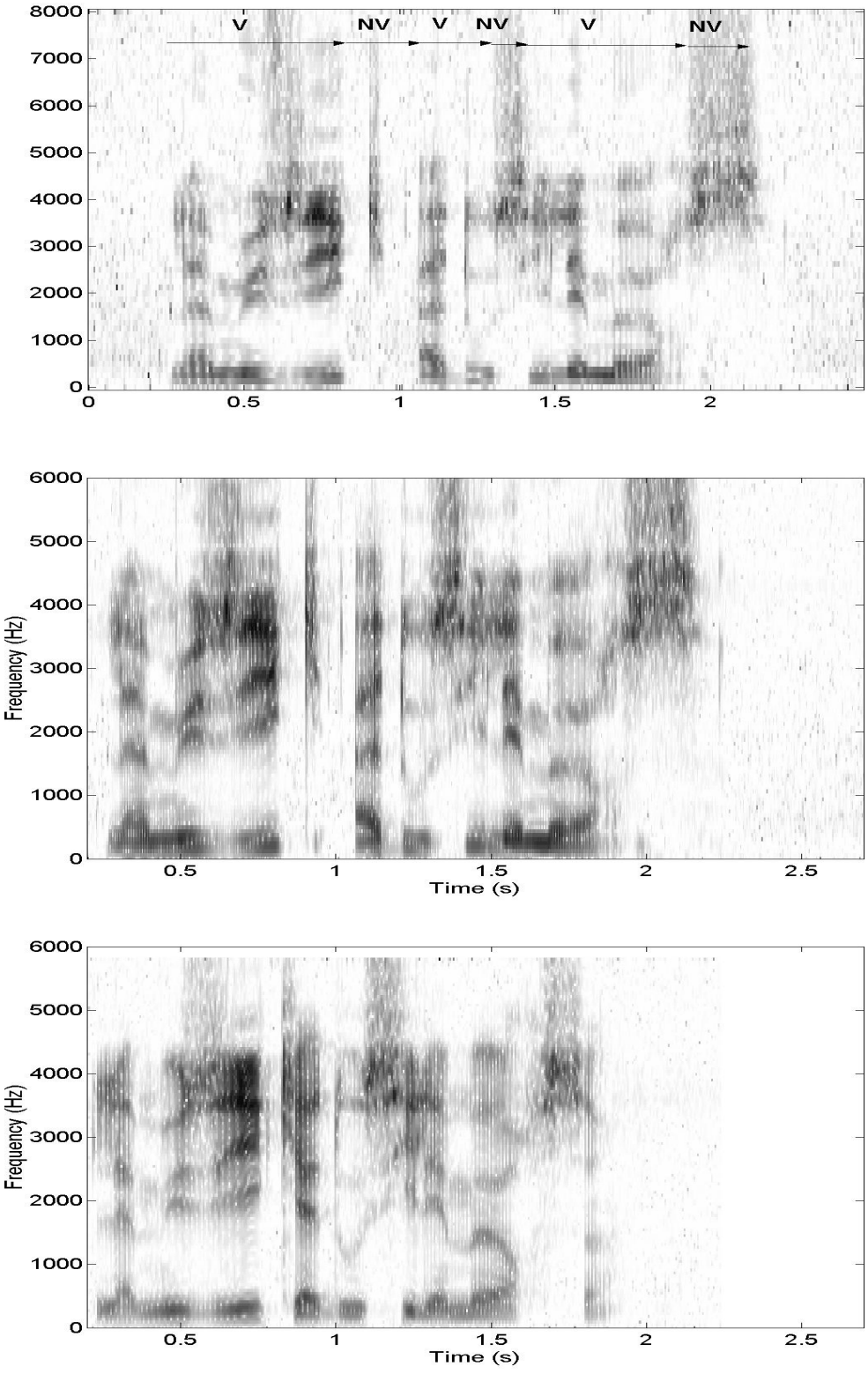


FIGURE III.1 – La phrase *"La musique adoucit les moeurs"* prononcée trois fois par le même locuteur

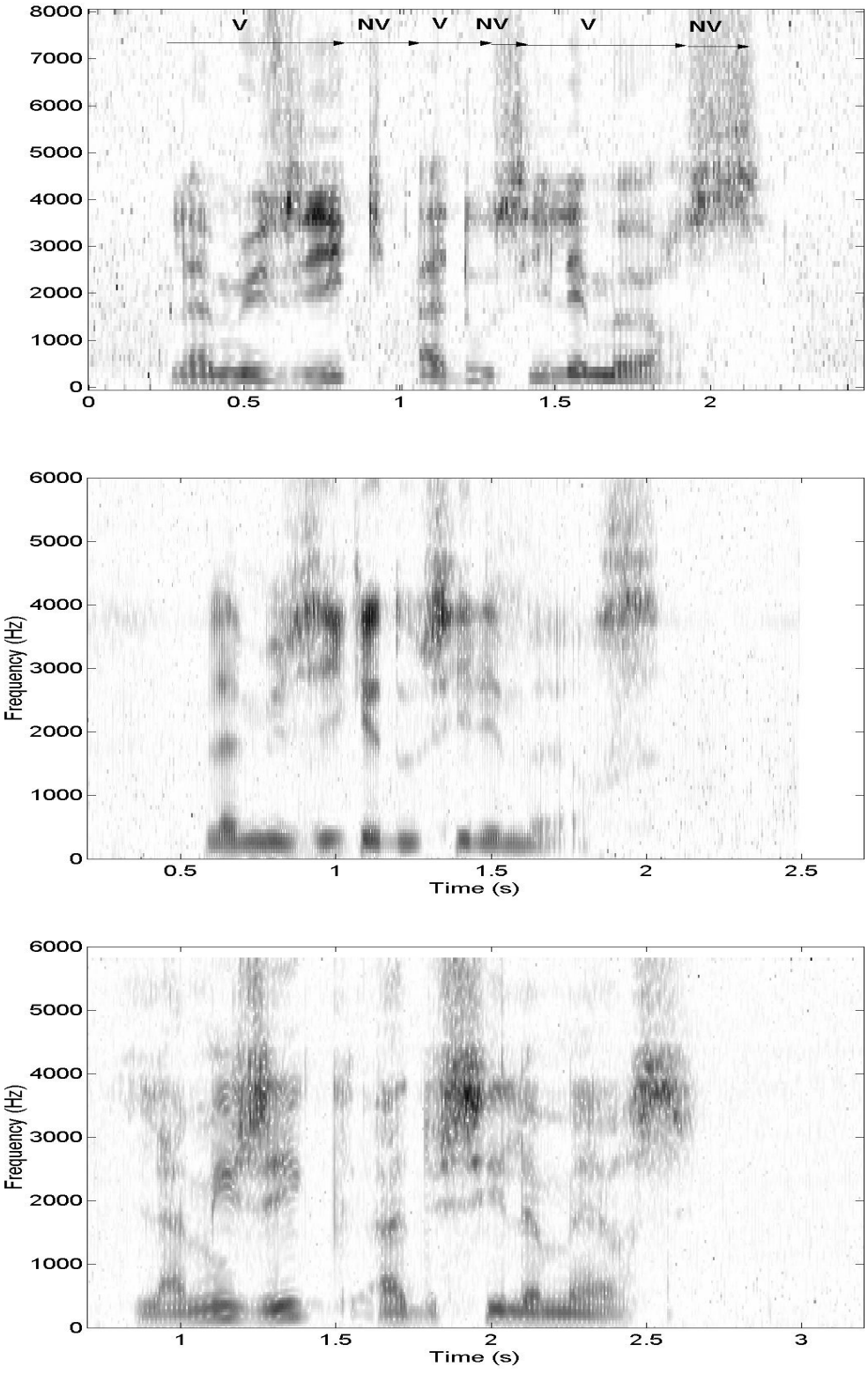


FIGURE III.2 – La phrase *"La musique adoucit les moeurs"* prononcée par trois locuteurs différents

environnement sans bruit de fond ni réverbération et avec un microphone de bonne qualité situé à une distance stable de la bouche. Bien évidemment, en pratique ces conditions ne sont pas souvent réunies, et on parle ainsi souvent de reconnaissance en *conditions difficiles*. De nombreuses variables perturbant les performances des systèmes de reconnaissance ont ainsi été identifiées ([5],[13]) :

- *les bruits d'environnement* : tels que les bruits additifs stationnaires (bruit de fond,...) ou non stationnaires (bruit de porte, sonneries de téléphone etc....)
- *Les déformations acoustiques* : telles que les distorsions non-linéaires dues à la qualité et dynamique variables des microphones et dues aux effets de réverbération dans une pièce
- *La largeur de bande du signal de parole* : (par exemple pour les applications téléphoniques la bande passante sera naturellement limitée entre 300 et 3400 Hz)
- *Les variations d'élocution* : ou élocution altérée comprenant entre autres l'effet Lombard [1], le stress physique ou émotionnel, une vitesse d'élocution inhabituelle, des hésitations ("euh...ben") ainsi que de divers bruits de production (bruits de bouche ou de respiration).

III.2 Approches pour la reconnaissance de parole

La reconnaissance de la parole consiste à extraire l'information lexicale contenue dans un signal acoustique (signal électrique obtenu à la sortie d'un microphone). D'une façon générale, on peut distinguer trois principales familles de méthodes pour la reconnaissance de la parole :

- *Les approches basées sur les connaissances* qui consistent à utiliser les connaissances phonétiques.
- *Les approches statistiques* de reconnaissance des formes qui consistent à apprendre une segmentation et une classification par apprentissage sur des données puis à utiliser cette classification pour la reconnaissance. Ce sont actuellement les approches les plus utilisées en reconnaissance de la parole.
- *Les approches d'intelligence artificielles* sont des approches hybrides qui incluent les approches à base de systèmes experts. Nous ne décrirons pas cette approche dans ce cours car elle est maintenant très peu utilisée même si certains concepts permettent de montrer l'intérêt des réseaux de neurones pour certaines phases de la reconnaissance.

III.2.1 Les approches statistiques

Ces approches utilisent directement la parole sans effectuer une détermination explicite des caractéristiques (au sens phonético-acoustique) ou de segmentation explicite. Les méthodes employées ont deux phases principales :

- *L'apprentissage* des unités élémentaires (ou patterns) vocales (ces unités ou segments peuvent être un son, un mot, une phrase etc..) La connaissance de la parole est apportée au système à travers cette phase d'apprentissage. Le concept de base est qu'un nombre suffisamment grand de chaque unité est inclus dans l'ensemble d'apprentissage et que la procédure d'apprentissage est capable de caractériser les propriétés acoustiques de chaque unité.
- *La reconnaissance* qui permet de reconnaître une unité par comparaison. Dans cette étape, une comparaison directe entre le signal de parole à reconnaître avec chaque unité élémentaire apprise durant la phase d'apprentissage permet de classifier le signal d'entrée en fonction de ces unités.

1. L'effet Lombard regroupe toutes les modifications (pas toujours audibles) du signal acoustique lors d'une élocution en milieu bruité.

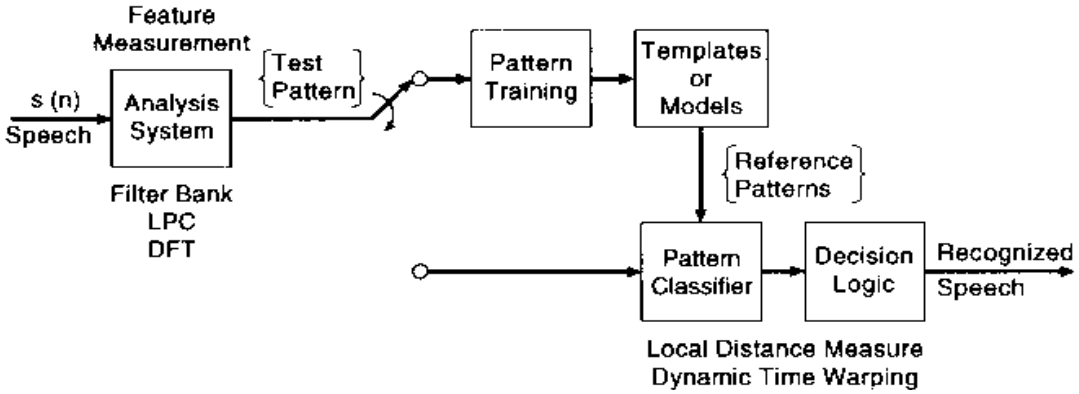


FIGURE III.3 – Schéma bloc d'un système de reconnaissance vocale par une approche statistique [22]

Cette approche est actuellement la plus répandue en reconnaissance de la parole et sera donc plus développée dans ce cours. On donne souvent trois raisons principales pour expliquer le succès de ces approches :

- *La simplicité de mise en oeuvre* : ces méthodes sont accessibles et reposent sur des bases mathématiques rigoureuses et sur la théorie de l'information.
- *La robustesse et l'invariance de l'approche* aux différents vocabulaires, utilisateurs, unités choisies,etc...Ainsi, cette approche est applicable pour une très large classe d'unités de parole (phonèmes, mots, phrase, etc...), d'environnements, de conditions de transmission etc....
- *Les performances :* cette famille d'approches permet d'obtenir d'excellents résultats qui ont été maintes fois démontrés.

On donne sur la figure III.3 un schéma bloc d'un système de reconnaissance de la parole par une approche statistique. On voit sur ce diagramme qu'une telle approche possède quatre étapes principales :

1. *L'extraction de paramètres* (dénommée *feature Measurement* sur la figure) qui contient un module de traitement du signal et d'analyse acoustique transformant le signal de parole en une séquence de vecteurs acoustiques (*Test pattern*). Ces vecteurs sont usuellement obtenus à l'aide d'une analyse spectrale utilisant des bancs de filtres, une analyse par prédiction linéaire (LPC) ou une transformée de Fourier discrète.
2. Une étape *d'apprentissage* durant laquelle plusieurs vecteurs acoustiques correspondant aux sons d'une même classe sont utilisés pour créer un *représentant* caractéristique de cette classe. Ce représentant peut être un vecteur (obtenu par moyennage par exemple) ou être un modèle qui caractérise les statistiques des paramètres de ce représentant.
3. Une étape *de classification* : dans laquelle le vecteur acoustique inconnu est comparé aux représentants de chaque classe à l'aide d'une mesure de similarité (distance). Cette distance doit tenir compte d'un désalignement temporel en raison des différences de vitesse d'élocution (appelé *Dynamic Time Warping* en anglais)
4. Une *étape de décision* qui sélectionne le meilleur représentant.

Le choix de la paramétrisation acoustique, des modèles et de la classification utilisés sont les principales différences que l'on pourra trouver entre les différents systèmes.

Dans ce cours, on verra certains aspects de ces approches. Faisons, dès maintenant, quelques remarques :

- Les performances du système sont dépendantes des données utilisées et notamment de l'importance (en taille) de ces données. En général, plus on dispose de données, plus le système de reconnaissance sera performant. On comprend ici l'importance des bases de données. Un exemple de telle base est donné à la section **??**.
- Assez peu de connaissances directement liées au signal de parole sont utilisées explicitement. Ainsi, ces approches seront relativement insensibles au choix des mots de vocabulaire, de la tâche, de la syntaxe, etc...
- Les contraintes en coût calcul peuvent devenir importantes sachant que les procédure d'apprentissage et de reconnaissance sont en gros linéairement proportionnelles au nombre d'unités à reconnaître.

III.3 Paramétrisation

Nous verrons dans cette partie comment est réalisée la paramétrisation du signal de parole en vue de sa reconnaissance. Cette paramétrisation est réalisée par un module de traitement du signal (dénommé *Acoustic Front-end*) en raison de sa position dans la chaîne générale d'un système de reconnaissance.

Il réalise une analyse spectrale du signal. Cette analyse est généralement faite suivant l'une des méthodes suivantes ([13]) :

- Par banc de filtres (typiquement entre 10 à 30 bandes fréquentielles)
- Par transformée de Fourier (FFT), cette méthode étant bien évidemment un cas particulier de la précédente
- Une approche basée sur les coefficients cepstraux, ces derniers ayant pu être calculés à partir de la sortie d'un banc de filtres
- en dérivant une enveloppe spectrale à partir d'une analyse par prédiction linéaire (LPC).

Les méthodes par bancs de filtres ont été très utilisées mais ont tendance à être maintenant remplacées par des approches plus spécifiques.

L'approche par Transformée de Fourier rapide est très souvent préférée en raison de sa simplicité mais aussi bien sur en raison des algorithmes de calcul rapide qui existent. Les valeurs spectrales obtenues à des intervalles égaux sont souvent ré-échantillonnés sur une échelle logarithmique. Dans un souci de prendre encore plus en compte les caractéristiques de l'audition, d'autres échelles plus appropriées sont couramment utilisées. Il s'agit de l'échelle Bark et de l'échelle Mel (de loin la plus utilisée en reconnaissance de la parole).

L'échelle Bark est basée sur les bandes critiques telles qu'elles sont perçues par l'oreille. Les valeurs de l'échelle Bark sont représentées dans le tableau figure III.4 et sont assez proches des valeurs prises sur une échelle logarithmique. Il existe plusieurs formules analytiques pour approcher la relation qui existe entre les fréquences f et les nombres en bande critiques z exprimés en Bark. La formule analytique suivante possède l'avantage de proposer une formule inversible ([14]) sachant que des facteurs de correction sont appliqués pour les valeurs en dessous de 2 Bark et les valeurs au dessus de 20.1 Bark :

Bark	Lower (Hz)	Center (Hz)	Upper (Hz)	Bark	Lower (Hz)	Center (Hz)	Upper (Hz)
0-1	0	50	100	**12-13**	1720	1850	2000
1-2	100	150	200	**13-14**	2000	2150	2320
2-3	200	250	300	**14-15**	2320	2500	2700
3-4	300	350	400	**15-16**	2700	2900	3150
4-5	400	450	510	**16-17**	3150	3400	3700
5-6	510	570	630	**17-18**	3700	4000	4400
6-7	630	700	770	**18-19**	4400	4800	5300
7-8	770	840	920	**19-20**	5300	5800	6400
8-9	920	1000	1080	**20-21**	6400	7000	7700
9-10	1080	1170	1270	**21-22**	7700	8500	9500
10-11	1270	1370	1480	**22-23**	9500	10500	12000
11-12	1480	1600	1720	**23-24**	12000	13500	15500

FIGURE III.4 – Tableau récapitulant les valeurs de l'échelle Bark [14]). Notons que sur l'échelle Bark, les valeurs entières correspondent aux limites de l'intervalle. Ainsi, 8 Bark correspond à 920 Hz, et 1000 Hz correspond à 8.5 Bark.

$$z' = \frac{26.81 f}{(1960+f)} - 0.53 \tag{III.1}$$

(III.2)

$$\text{si} \quad z' < 2.0 \quad \text{Bark}, \quad z = z' + 0.15(2.0 - z') \tag{III.3}$$

$$\text{si} \quad z' > 20.1 \quad \text{Bark}, \quad z = z' + 0.22(z' - 20.1) \tag{III.4}$$

La formule inverse est alors donnée par l'équation III.5 :

$$\text{si} \quad z < 2.0 \quad \text{Bark}, \quad \text{alors} \quad z' = 2.0 + \frac{(z-2.0)}{0.85} \tag{III.5}$$

$$\text{si} \quad z > 20.1 \quad \text{Bark}, \quad \text{alors} z' = 20.1 + \frac{(z-20.1)}{1.22} \tag{III.6}$$

$$\text{sinon} \quad z' = z \tag{III.7}$$

(III.8)

$$\text{et} \quad f = 1960 \frac{(z'+0.53)}{(26.28 - z')} \tag{III.9}$$

L'échelle Mel correspond à une approximation de la sensation psychologique de hauteur d'un son. De même que pour les formules analytiques de l'échelle Bark, il n'existe pas d'échelle Mel unique. Une relation couramment utilisée reliant la fréquence f et l'échelle Mel, $mel(f)$, est donnée dans ([13]) :

$$mel(f) = 1000 \log_2(1 + \frac{f}{1000}) \tag{III.10}$$

Notons que la fréquence 1000 Hz correspond à la valeur 1000 mel.

L'utilisation de l'échelle Mel conduit à l'une des paramétrisations les plus utilisées en reconnaissance de la parole : les coefficients *MFCC* (pour *Mel Frequency Cepstral Coefficients*) qui sont décrits au paragraphe III.3.3

III.3.1 Représentation cepstrale

Comme nous l'avons vu précédemment, la parole peut être représentée sous la forme d'un modèle source-filtre. Cette représentation permet ainsi de représenter le signal de parole $s(t)$ sous la forme du convolution du signal source $g(t)$ par la réponse impulsionnelle du filtre $h(t)$ représentant le conduit vocal :

$$s(t) = g(t) * h(t) \tag{III.11}$$

L'étude de ce signal à l'aide de la FFT présente un défaut particulier liée à cette convolution qui rend difficile l'observation de la seule contribution du conduit vocal. Le cepstre (parfois appelé lissage cepstral) permet de séparer les contributions respectives de la source et du conduit vocal.

En effet, l'équation III.11 se récrit dans le domaine spectral sous la forme :

$$S(\omega) = G(\omega)H(\omega) \tag{III.12}$$

où $S(\omega)$,$G(\omega)$ et $H(\omega)$ représentent respectivement les transformées de Fourier de $s(t)$,$g(t)$ et $h(t)$.

Le cepstre qui est défini par le logarithme de la transformée de Fourier inverse du module de $S(\omega)$ s'écrit donc sous la forme :

$$c(\tau) = FFT^{-1} \log |S(\omega)| = FFT^{-1} \log |G(\omega)| + FFT^{-1} \log |H(\omega)| \tag{III.13}$$

On peut alors noter que le spectre s'exprime comme la somme de deux termes. Le premier terme $FFT^{-1} \log |G(\omega)|$ est caractéristique de la source et représente ainsi la structure fine, tandis que le second terme est caractéristique de l'enveloppe spectrale et représente la contribution du conduit vocal. Le paramètre τ homogène à un temps est appelé *quéfrence*. A l'aide de cette représentation, il est possible d'isoler soit le pic (qui correspond au pitch) qui se trouve dans la région des hautes quéfrences (on a ici une méthode d'estimation de la fréquence fondamentale) soit d'isoler la partie correspondant aux basses quéfrences qui représente une version lissée de l'enveloppe spectrale. Ce procédé de séparation des éléments cepstraux est appelé un *liftrage* (par dérivation de l'appellation filtrage). On donne figure III.5 un exemple de plusieurs liftres permettant de séparer les contributions source et conduit vocal (d'après [7]).

Lorsque le cepstre est obtenu en calculant la transformée de Fourier discrète, on obtient la forme suivante :

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{2j(\pi)kn/N} \qquad \text{pour} \qquad 0 \leq n \leq N-1 \tag{III.14}$$

La figure III.6 donne un exemple de cepstre (d'après [13]).

III.3.2 La paramétrisation MFCC

Si de nombreuses paramétrisations sont possibles pour la reconnaissance de parole, il existe trois représentations qui ont été plus particulièrement étudiées et qui se retrouvent dans la grande majorité des systèmes actuels de reconnaissance vocale :

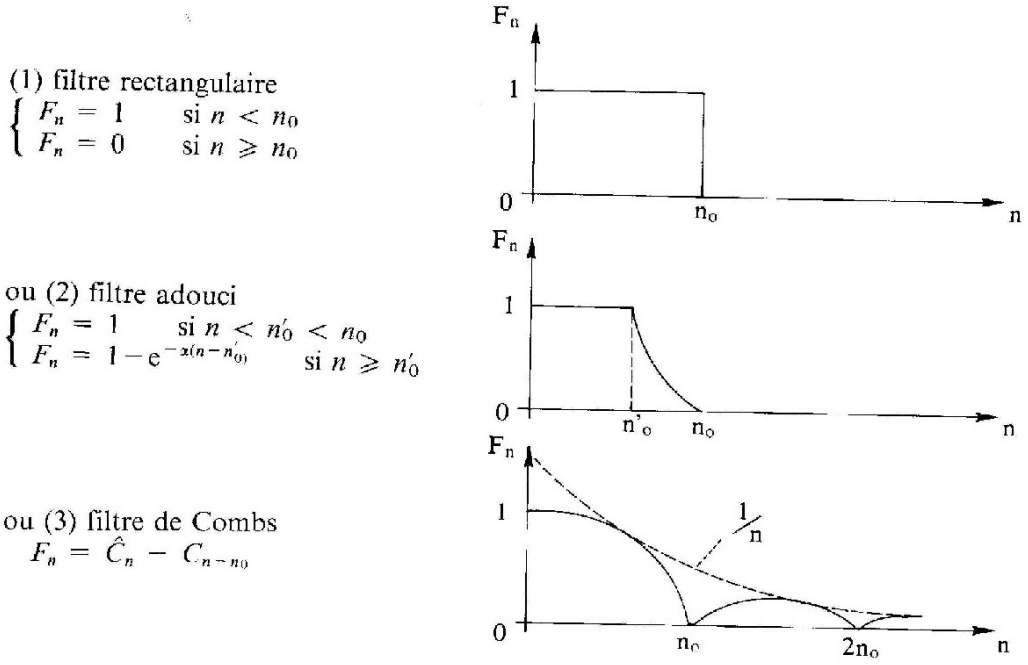


FIGURE III.5 – Exemples de liftres (d'après [7])

- La représentation cepstrale à base de prédiction linéaire : les paramètres LPCC
- La représentation à base de prédiction linéaire perceptuelle : les paramètres PLP
- La représentation cesptrale utilisant des bancs de filtres sur une échelle Mel : les paramètres MFCC

Nous ne décrirons ci-dessous que la paramétrisation MFCC laissant les autres paramétrisations pour un cours plus avancé.

III.3.3 La paramétrisation MFCC

La paramétrisation MFCC (Mel-Frequency Cepstral coefficients) est probablement la paramétrisation la plus répandue dans les systèmes de reconnaissance actuels. Nous donnons ci-dessous les principales étapes de cette paramétrisation :

1. *Fenêtrage du signal* Le signal de parole est séparé en trames de N échantillons, chaque trame étant séparée de M échantillons. Dans le cas courant où $M < N$ on dira qu'il y a recouvrement (*overlap* en anglais) entre les trames. Un exemple est donné sur la figure III.7 pour $M = \frac{1}{3}N$. En pratique, la longueur N d'une trame est couramment choisie de façon à avoir des trames dont la durée est de l'ordre de 20ms associé à un recouvrement entre trames de 50% correspondant à une valeur de $M = \frac{N}{2}$. L'opération précédente consiste ainsi à appliquer une fenêtre rectangulaire de durée finie sur l'ensemble du signal. Pour réduire les effets dus aux discontinuités aux bords de la fenêtre, il est fréquent de pondérer une trame de longueur N par une fenêtre de pondération. L'une des fenêtres les plus utilisée est la *fenêtre de Hamming*. Cette opération donne la trame fenêtrée :

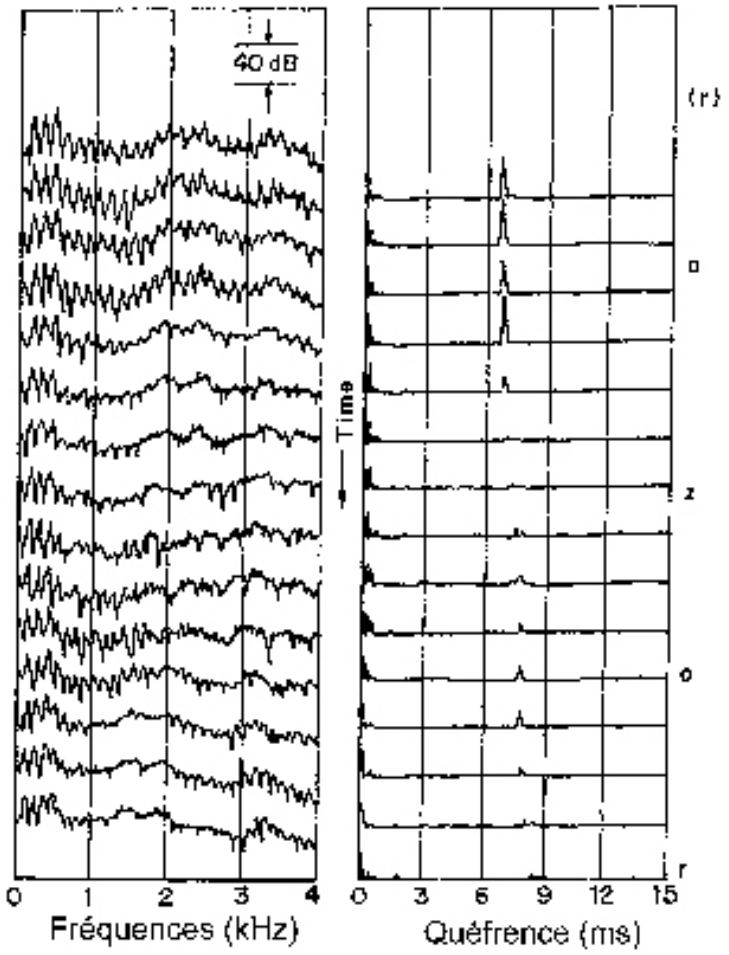


FIGURE III.6 – Exemple de spectres à court terme (gauche) et de cepstre (droite) pour une voix d'homme prononçant le mot "razor" (d'après [13])

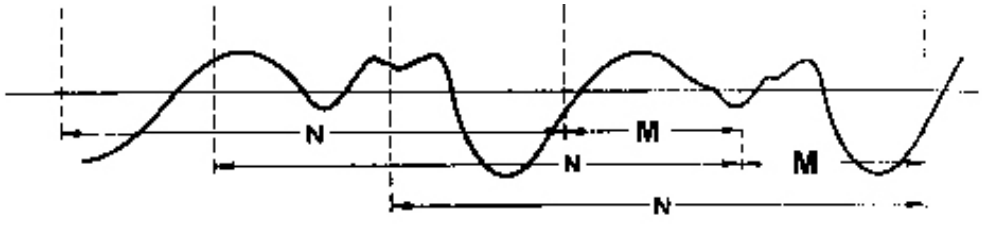


FIGURE III.7 – Exemple de recouvrement (*overlap* en anglais) entre trames pour $M = \frac{1}{3}N$. (d'après [22])

$$s_w(n) = \tilde{s}(n)w(n) \quad \text{(III.15)}$$

$$\text{où } w(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) \text{ avec } 0 \leq n \leq N-1 \quad \text{(III.16)}$$

2. *Calcul de la transformée de Fourier rapide* (FFT)pour chaque trame du signal de parole
3. *Filtrage par un banc de filtres MEL.* Cette opération permet d'obtenir à partir du spectre $S(k)$ de chaque trame, un spectre modifié qui est en fait une suite de coefficients, noté $\tilde{S}(k)$, représentant l'énergie dans chaque bande fréquentielle k (définies sur l'échelle Mel), pour $k = 1 \ldots K$. En pratique, on utilise des filtres triangulaires de largeur de bande constante et régulièrement espacées sur l'échelle Mel (On peut par exemple choisir un espacement entre filtres de 150 mels et une largeur des filtres triangulaire prise à leur base de 300 mels).
4. *Calcul des coefficients MFCC* : Les coefficients MFCC sont alors obtenus en effectuant une transformée en cosinus discrête inverse (de type II) du logarithme des coefficients $\tilde{S}(k)$:

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n(k - \frac{1}{2})\frac{\pi}{K}\right] \qquad \text{pour} \quad n = 1, 2, \ldots, L \quad \text{(III.17)}$$

   où L est le nombre de coefficients cepstraux désirés.
5. *Pondération* : En raison de la grande sensibilité des premiers coefficients cepstraux sur la pente spectrale générale et de la sensibilité au bruit des coefficients cepstraux d'ordre élevé, il est courant de pondérer ces coefficients pour minimiser cette sensibilité. Cette pondération pourra s'écrire sous la forme :

$$\hat{c}_m = w(m)c_m \qquad \text{pour} \quad 1 \leq m \leq Q \quad \text{(III.18)}$$

   où Q est le nombre de coefficients cepstraux.
   La fenêtre de pondération cepstrale est en fait un liftre passe bande dont un choix approprié peut être :

$$w(m) = \left[1 + \frac{Q}{2} sin(\frac{\pi m}{Q})\right] \qquad \text{pour} \quad 1 \leq m \leq Q \quad \text{(III.19)}$$

   Cette fenêtre tronque le nombre de coefficients et diminue le poids des premiers et derniers coefficients.
6. *Calcul des dérivées temporelles* Δ, Δ^2 *:* La représentation cepstrale donne une bonne représentation des propriétés fréquentielles locales du signal (i.e. pour une fenêtre de signal donnée). Une représentation améliorée peut être obtenue en incluant de l'information liée à l'évolution temporelle des coefficients cepstraux. Celle ci peut être obtenue par exemple à l'aide des dérivées premières et secondes des coefficients cepstraux. Soit $c_m(t)$ les coefficients cepstraux obtenus à l'instant t (ou plus précisément à la fenêtre d'indice t). Cette suite est obtenue à des instants discrets et ainsi il est bien connu qu'un simple moyennage aux différences ne permet pas d'obtenir des estimations non bruitées. Ainsi, la dérivée est souvent obtenue en effectuant une moyenne sur un plus grand horizon temporelle sous la forme :

$$\Delta c_m(t) \approx \mu \sum_{k=-K}^{K} k c_m(t+k) \quad \text{(III.20)}$$

   où μ est une constante de normalisation et $(2K+1)$ est le nombre de trames utilisées pour ce calcul.

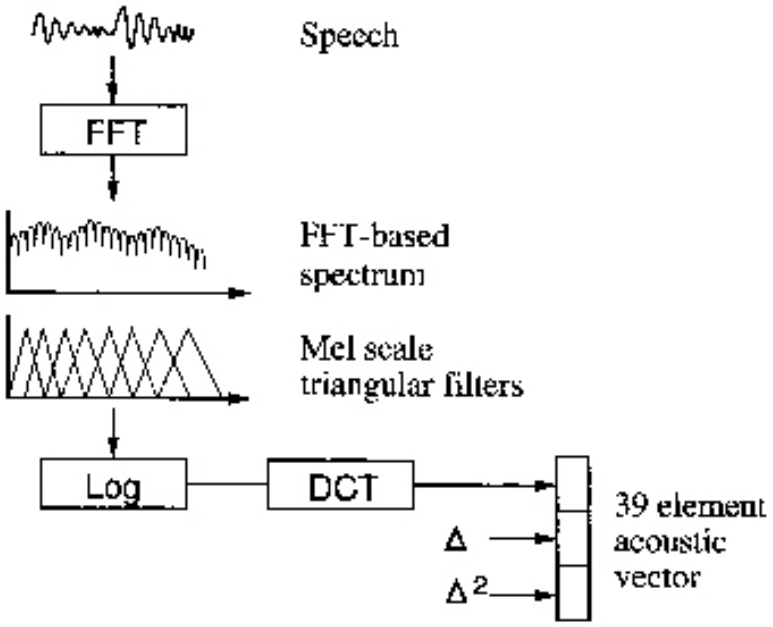


FIGURE III.8 – Schéma bloc de la paramétrisation MFCC (*MFCC-based Front-End processor* d'après [13])

Une implémentation classique de la paramétrisation MFCC consiste à prendre les 13 premiers coefficients cepstraux (en omettant l'énergie représentée par c_0) et à construire des vecteurs acoustiques de 39 éléments incluant les dérivées première (Δ) et seconde (Δ^2) de ces coefficients. La figure III.8 donne un schéma bloc de cette implémentation classique.

III.4 Distances et mesures de distorsion spectrale

Un des points clés en reconnaissance de la parole est lié à la façon dont les segments de parole (ou leur représentation paramétrique) vont être comparés pour déterminer leur similarité (ou de façon équivalente leur distance). Il existe un nombre important de techniques permettant une telle comparison, ces techniques étant bien évidemment dépendantes de la paramétrisation utilisée. Nous allons voir ci-dessous quelques unes des distances les plus utilisées en reconnaissance vocale.

III.4.1 Distance : aspects mathématiques et perceptuels

La mesure de similarité de deux segments de parole représentés par leurs vecteurs acoustiques peut être effectuée de manière rigoureuse. Soit, $\mathbf{x}$ et $\mathbf{y}$ deux vecteurs acoustiques définis dans un espace vectoriel χ. On peut définir une distance d sur cette espace comme étant une fonction à valeurs réelles telle que :

- $0 \leq d(\mathbf{x},\mathbf{y}) < \infty$ pour $\mathbf{x},\mathbf{y} \in \chi$ et $d(\mathbf{x},\mathbf{y}) = 0$ si et seulement $\mathbf{x} = \mathbf{y}$
- $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$ pour $\mathbf{x},\mathbf{y} \in \chi$
- $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{y},\mathbf{z})$ pour $\mathbf{x},\mathbf{y} \in \chi$

On dira de plus que la distance est invariante si :

- $d(\mathbf{x}+\mathbf{z},\mathbf{y}+\mathbf{z}) = d(\mathbf{x},\mathbf{y}$

Les 3 premières propriétés font référence au fait qu'une distance est "définie positive". Si seules ces propriétés sont vérifiées, on parlera de *mesure de distortion*.

En traitement de la parole, il est important également d'avoir une distance qui prennent en compte des aspects perceptuels. Intuitivement, on comprend que des spectres différents vont pouvoir donner lieu à une grande distance et qui pourtant pourront pourraient être très proches au niveau perceptuel. Par exemple, un certain nombre de changements spectraux ne changent pas le son (i.e. le phonème) perçu. Ces changements incluent :

- La *pente spectrale* (ou spectral tilt) : $S'(\omega) = S(\omega) \cdot \omega^{\alpha}$ où α est un facteur de pente spectrale
- *Filtrage passe-bas* ou passe haut à condition que les fréquences de coupures soient suffisament basses ou suffisamment hautes.
- *Filtrage notch*,où $S'(\omega) = S(\omega)|H_N(e^{j\omega})|^2$ où $H_N(e^{j\omega})$ est un filtre passe-tout excepté sur une bande très étroite en fréquence où le signal sera fortement atténué.

Par contre, certains changements spectraux auront un impact direct sur le son (i.e. phonème) perçu, comme par exemple :

- Les déplacements de position des formants,
- Les changements de largeur de bande de ces formants

Si de nombreuses études ont vus le jour pour définir des distances psychoacoustiques (l'une d'entre elles consiste à étudier les plus petites différences perceptibles pour un certain nombre de paramètres (fréquence fondamentale, la position et largeur de bande des formants, etc. voir [**?**] par exemple), l'utilisation de telles distances s'avère difficile en pratique. Il est ainsi souvent préféré en reconnaissance de la parole d'utiliser des distances ou mesures de distorsion définies rigoureusement tout en intégrant le fait que ces mesures doivent être en accord avec les aspects perceptifs importants en parole. Les mesures de distorsion spectrale entrent dans ce cadre puisqu'elles sont définies rigoureusement et que les études psychoacoustiques montrent quasiment toutes que les différences perçues peuvent être interprétées en termes de différences spectrales.

III.4.2 Distance Log-spectrale

Les distances Log-spectrales sont des mesures de distorsion particulièrement utiles et sont réellement appropriées sur un point de vue perceptuel. Soit $S(\omega)$ et $S'(\omega)$ deux spectres dont nous voulons calculer la différence. Un choix naturel de mesure de distorsion entre S et S' est l'ensemble des normes L_p définies par :

$$d(S, S')^p = \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^p \frac{d\omega}{2\pi} \tag{III.21}$$

La figure III.9 représente la différence spectrale logarithmique calculée à partir des transformées de Fourier de deux signaux $s(n)$ et $s'(n)$. On peut remarquer que cette différence est très bruitée (ou irrégulière). Pr ailleurs, on peut remarquer qu'une partie importante de ces irrégularités provient d'une différence de fréquence fondamentale qui n'est pas un paramètre importante pour l'identification phonétique (tout au moins pour les langues qui ne sont pas "à tons" tels que le chinois). On peut alors utiliser cette norme L_p sur les modèles tout pôle d'une prédiction linéaire qui sera alors définie par :

$$d_{lpc}(S, S')^p = \int_{-\pi}^{\pi} |\log \frac{\sigma^2}{|A(e^{j\omega})|^2} - \log \frac{\sigma'^2}{|A'(e^{j\omega})|^2}|^p \frac{d\omega}{2\pi} \tag{III.22}$$

La figure III.10 représente la différence obtenue à partir des modèles de prédiction linéaire des signaux $s(n)$ et $s'(n)$ et on constate en comparant cette figure à la figure III.9 que la différence est

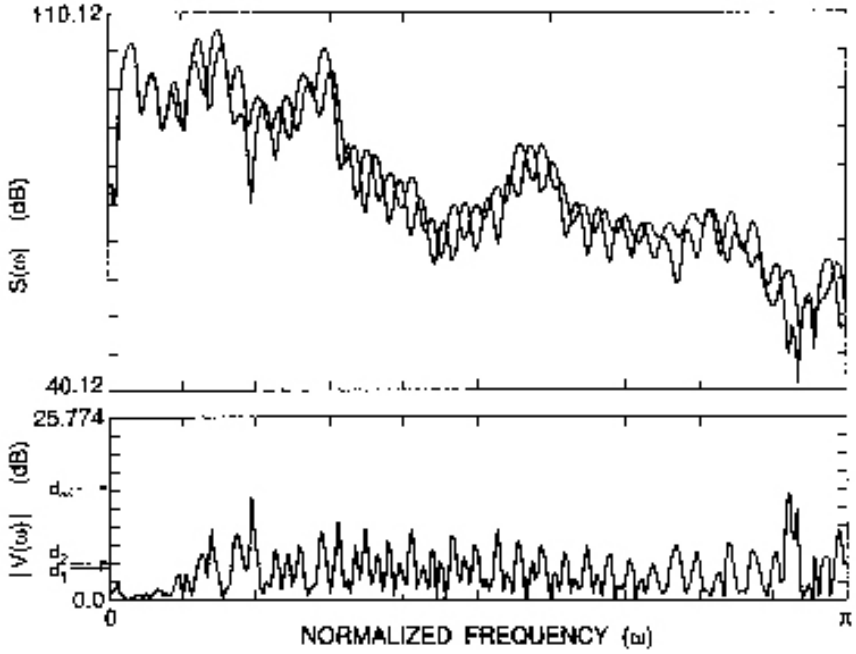


FIGURE III.9 – Spectre d'amplitude $S(\omega)$ et $S'(\omega)$ (en haut) et le module de leur différence logarithmique (en bas) (d'après [22])

beaucoup plus régulière et est, bien sur, moins sensible aux différences de fréquence fondamentale.

III.4.3 Distances cepstrales

Sachant que la paramétrisation cepstrale est l'une des plus utilisée en reconnaissance vocale, il est appréciable de disposer de distances cepstrales. En utilisant le théorème de Parseval, il est possible de relier la distance cepstrale d_2 à la distance spectrale logarithmique L_2 sous la forme :

$$d_2^2 = \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 \frac{d\omega}{2\pi} \quad \text{(III.23)}$$

$$= \sum_{n=-\infty}^{\infty} (c_n - c_n^{'})^2 \quad \text{(III.24)}$$

où c_n et $c_n^{'}$ sont les coefficients cepstraux de $S(\omega)$ et $S'(\omega)$. En pratique, il n'est pas nécessaire de calculer cette somme pour un nombre infini de termes. La somme est ainsi tronquée à un nombre limité L de termes (où L est de l'ordre de 10 à 20) :

$$d_2^2 = \sum_{n=1}^{L} (c_n - c_n^{'})^2 \quad \text{(III.25)}$$

On peut étendre la distance précédente en intégrant une pondération permettant de diminuer la sensibilité au canal de transmission et à la variabilité inter-locuteur. Il est connu que la variabilité des premiers coefficients cepstraux est principalement due aux variations du canal de

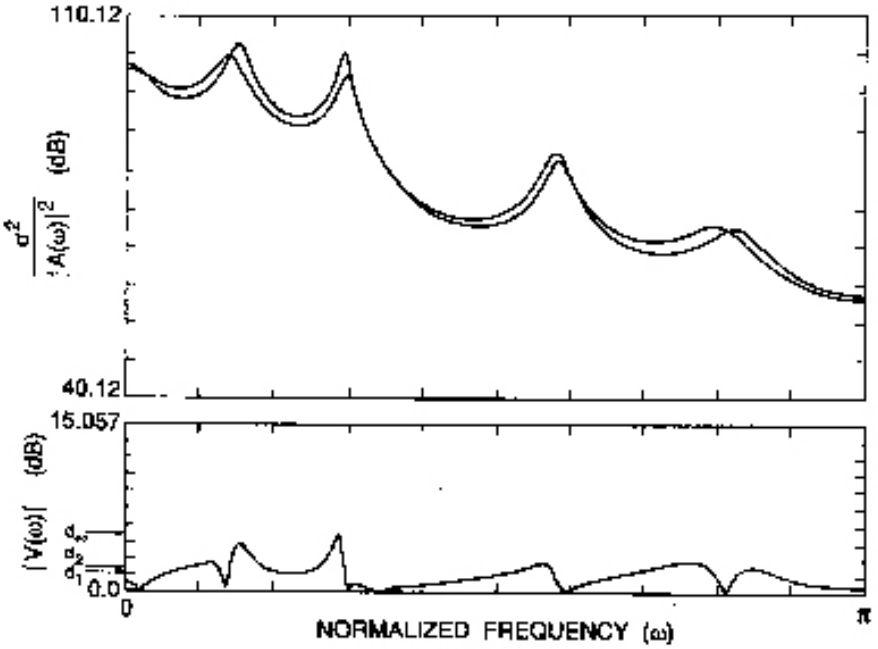


FIGURE III.10 – Modèles LPC $\frac{\sigma^2}{|A(e^{j\omega})|^2}$ et $\frac{\sigma'^2}{|A'(e^{j\omega})|^2}$ (en haut) et le module de leur différence logarithmique (en bas) (d'après [22])

transmission, aux caractéristiques du locuteurs et d'autres facteurs tels que l'effort vocal. Dans une optique de reconnaissance phonétique, il apparaît important de diminuer ainsi le poids de ces premiers coefficients dans le calcul de la distance. Ceci peut être réalisé à l'aide d'une fenêtre de pondération comme celle donnée par l'équation III.19. Dans ce cas, la distance s'écrira :

$$d_w^2 = \sum_{n=1}^{L} (w(n)c_n - w(n)c_n^{'})^2 \tag{III.26}$$

III.4.4 Distances cepstrales intégrant les Δ-cepstres

Il est également courant d'intégrer les dérivées première et seconde des coefficients cepstraux dans le calcul des distances cepstrales. On définit ainsi une différence cepstrale différentielle $d_{2\Delta}^2$ qui est proche de la distance spectrale différentielle :

$$d_{2\Delta}^2 = \int_{-\pi}^{\pi} |\frac{\partial \log S(\omega,t)}{\partial t} - \frac{\partial \log S'(\omega,t)}{\partial t}|^2 \frac{d\omega}{2\pi} \tag{III.27}$$

$$\simeq \sum_{n=-\infty}^{\infty} (\Delta c_n - \Delta c_n^{'})^2 \tag{III.28}$$

De même, pour les dérivées secondes on aura :

$$d^2_{2\Delta(2)} = \int_{-\pi}^{\pi} |\frac{\partial^2 \log S(\omega,t)}{\partial^2 t} - \frac{\partial^2 \log S'(\omega,t)}{\partial^2 t}|^2 \frac{d\omega}{2\pi} \quad \text{(III.29)}$$

$$\simeq \sum_{n=-\infty}^{\infty} (\Delta c_n^{(2)} - \Delta c_n^{'(2)})^2 \quad \text{(III.30)}$$

Il est alors possible de combiner ces distances de manière assez simple avec la distance cepstrale pour donner :

$$d_{2,\Delta,\Delta^{(2)}} = \gamma_1 d_2^2 + \gamma_2 d_{2\Delta}^2 + \gamma_3 d_{2\Delta^{(2)}}^2 \quad \text{(III.31)}$$

où γ_1, γ_2 et γ_3 sont des poids utilisés pour ajuster la contribution de chaque distance. En pratique, on posera $\gamma_1 + \gamma_2 + \gamma_3 = 1$.

III.5 Alignement Temporel et Programmation dynamique

Nous avons vu dans les sections précédentes plusieurs approches pour la comparaison de spectres de parole sur la base d'un segment (une trame) de parole. Bien évidemment, cette comparaison doit être menée pour l'ensemble du mot ou de la phrase prononcée. Hors, cette comparaison est confrontée au fait que deux mots ou phrases sont très rarement prononcées avec la même vitesse d'élocution et ainsi les deux séquences X (entrée que l'on cherche à reconnaître) et Y^k (référence apprise) n'auront pas en général la même durée. La solution la plus simple sera alors d'effectuer une *déformation temporelle linéaire*, c'est à dire associer plusieurs vecteurs de référence à un vecteur d'entrée (ou vice-versa si le vecteur d'entrée est plus long que le vecteur de référence). Ainsi une déformation temporelle linéaire pourra s'écrire :

$$d(\chi, \xi) = \sum_{i_x=1}^{T_x} d(i_x, i_y) \quad \text{(III.32)}$$

où i_x et i_y vérifient la relation

$$i_y = \frac{T_y}{T_x} i_x \quad \text{(III.33)}$$

Une illustration de cet alignement temporel linéaire est donnée figure III.11

Cependant, cet alignement n'est pas optimal car il suppose que le mot d'entrée est prononcé entièrement plus rapidement (resp. plus lentement) et toujours dans la même proportion. En pratique, il est possible que certaines parties (phonèmes) soient prononcées plus rapidement sur le mot test que pour le mot de référence alors que d'autres sections seraient prononcées plus lentement. On peut ainsi définir un alignement temporel plus général qui est couramment appelé *Déformation Temporelle dynamique* (ou DTW pour *Dynamic Time Warping*). Cette déformation utilise deux fonctions de déformation ϕ_x et phi_y qui relient les indices des deux segments de parole (i_x et i_y respectivement) à un axe temporel commun k :

$$i_x = \phi_x(k) \text{ pour } k = 1, 2, \ldots, T \quad \text{(III.34)}$$

$$i_y = \phi_y(k) \text{ pour } k = 1, 2, \ldots, T \quad \text{(III.35)}$$

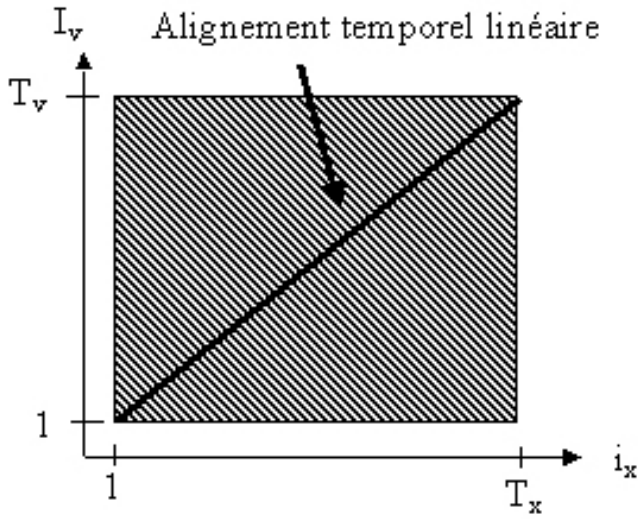


FIGURE III.11 – Alignement temporel linéaire

Il est ensuite possible de définir une mesure de similarité $d_\phi(\chi, \xi)$ à partir des fonctions de déformations sous la forme :

$$d_\phi(\chi, \xi) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)))m(k)/M_\phi \tag{III.36}$$

où $d(\phi_x(k), \phi_y(k)))$ mesure la distorsion spectrale pour les vecteurs $x_{\phi_x(k)}$ et $y_{\phi_y(k)}$, $m(k)$ est un coefficient (non-négatif) de pondération le long du chemin et M_ϕ est un facteur de normalisation.

La figure III.12 donne un exemple de normalisation temporelle dynamique.

Pour compléter la définition d'une mesure de similarité pour la paire $(\chi, \xi))$, il est nécessaire de spécifier un chemin ϕ. Ainsi, le problème est ramené à choisir un chemin de telle sorte que la mesure de similarité soit consistante. Un choix naturel (et populaire) est de définir $d(\phi_x(k), \phi_y(k)))$ comme étant le minimum de $d_\phi(\phi_x(k), \phi_y(k)))$ sur tous les chemins possibles, soit :

$$d(\chi, \xi)) = \min_\phi d_\phi(\chi, \xi) \tag{III.37}$$

III.5.1 Programmation dynamique

La programmation dynamique (ou Dynamic programmming) est une approche qui permet, sous certaines conditions, d'obtenir la solution optimale à un problème de minimisation d'un critère d'erreur sans devoir considérer toutes les solutions possibles ([3]).

Pour chercher la meilleure distance $D(T_x, T_y)$ entre deux séquences x et y, il suffit alors de chercher le chemin dans cette matrice D de façon à minimiser la somme des distances locales rencontrées pour aller d'un point initial (généralement (1,1) correspondant au début des mots test et référence) au point final (T_x, T_y) (correspondant à la fin des deux séquences).

La mise en oeuvre de cet algorithme se fait alors de manière très simple. La distance optimale est obtenue en calculant, pour chaque entrée (i_x, i_y), la distance cumulée $D(i_x, i_y)$ correspondant à la distance optimale que l'on obtient en comparant les deux sous-séquences (sous-politiques)

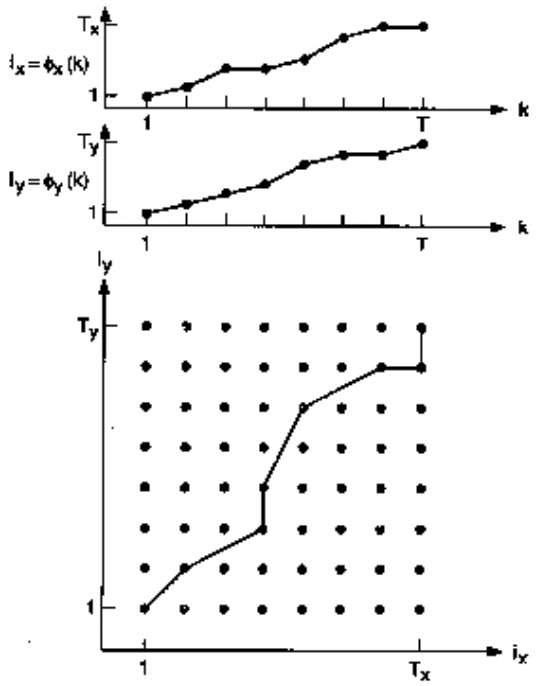


FIGURE III.12 – Exemple d'alignement dynamique (d'après [22]). La ligne en trait plein indique le chemin le long duquel la distance $d(\phi_x(k), \phi_y(k)))$ est évaluée.

correspondant aux i_x premiers vecteurs de test et aux i_y premiers vecteurs référence. La distance accumulée minimale sur le chemin entre $(1,1)$ et (i_x, i_y) sera ainsi donnée par :

$$D(i_x, i_y) = \min_{\phi_x, \phi_y, T'} \sum_{k=1}^{T'} d(\phi_x(k), \phi_y(k)) m(k) \tag{III.38}$$

où

$$\phi_x(T') = i_x \ \ ; \ \phi_y(T') = i_y \tag{III.39}$$

Notons que le coefficient de pondération M_ϕ a été ici omis puisqu'il ne dépend pas du chemin suivi et qu'il peut être déduit des contraintes. Il sera ainsi ré-injecté une fois que le point final aura été atteint. Ce facteur de normalisation est couramment pris comme la somme des poids le long du chemin choisi soit :

$$M_\phi = \sum_{k=1}^{T} m(k) \tag{III.40}$$

L'algorithme de programmation dynamique avec contraintes devient alors :

$$D(i_x, i_y) = \min_{(i'_x, i'_y)} [D(i'_x, i'_y) + \zeta((i'_x, i'_y), (i_x, i_y))] \tag{III.41}$$

où ζ est la distance pondéré entre le point (i'_x, i'_y) et le point (i_x, i_y) :

$$\zeta((i'_x, i'_y), (i_x, i_y)) = \sum_{l=0}^{L_s} d(\phi_x(T'-l), \phi_y(T'-l) m(T'-l) \tag{III.42}$$

où L_s est le nombre de déplacements dans le chemin pour aller de (i'_x, i'_y) à (i_x, i_y). Notons que :

$$\phi_x(T' - Ls) = i'_x \text{ et } \phi_y(T' - Ls) = i'_y \tag{III.43}$$

La figure III.13 donne un grand nombre de contraintes locales avec des pondérations associées qui ont été utilisées en reconnaissance vocale. Notons cependant que la contrainte la plus utilisée est aussi la plus simple (contrainte du haut sur la figure III.13)

Si la programmation dynamique est une technique utilisée dans de très nombreux domaines, son utilisation en reconnaissance vocale permet de définir des contraintes supplémentaires telles que :

- des *contraintes de monoticité* du chemin : le chemin commence au début des deux mots (point $(1, 1)$) et se termine à la fin des deux mots (point (T_x, T_y)).
- des *contraintes globales :* par exemple certaines contraintes permettant de réduire l'espace de recherche (en imposant que le chemin optimal reste dans une zone déterminée proche de la diagonale, voir figure III.14)
- des *contraintes locales :* les prédécesseurs sont limités à quelques éléments proches et garantissant un chemin strictement gauche droite (les phonèmes sont prononcés dans le même ordre dans le mot "test" et le mot "référence". On ajoutera comme il est montré figure III.13 des pénalités de transition ou poids suivant les chemins pris.

En pratique on peut résumer l'implémentation de la programmation dynamique sous la forme :

1. Initialiser la matrice D_A des distances cumulées avec la distance locale entre le premier vecteur de test et le premier vecteur de référence $D_A(1, 1) = d(1, 1)m(1)$ où $m(1) = 1$
2. Calculer les distance locales pour tous les autres éléments de la première colonne de D (soit $d(1, i)$ c'est à dire les distances entre le premier vecteur de test et tous les vecteurs de référence)
3. Si la transition verticale est autorisée, calculer à l'aide de l'équation (III.41)les distances accumulées $D_A(1, i)$ correspondant à la première colonne. Si la transition n'est pas autorisée, les distances accumulées de la première colonne est égale à l'infini (sauf bien entendu pour le point $(1, 1)$).
4. Passer à la colonne suivante, calculer les distances locales $d(2, i)$ et ensuite à l'aide de l'équation (III.41) calculer les distances accumulées $D(2, i)$ associées. Itérer sur toutes les colonnes.
5. Lorsque le dernier point est atteint, réinjecter le coefficient de normalisation $d(\chi, \xi) = \frac{D_A(T_x, T_y)}{M_\phi}$

Notons qu'après chaque itération, il n'est nécessaire de ne garder en mémoire que la dernière colonne de distances accumulées.

III.5.2 Reconnaissance de mots enchaînés à l'aide de la programmation dynamique

La reconnaissance de mots enchaînés est un problème plus complexe puisqu'il existe ici une co-articulation entre les mots et que les mots ne sont plus séparés par des silences. Comme il n'est pas envisageable de mettre en mémoire toutes les séquences de mots possibles, il va être nécessaire de segmenter (de façon automatique) la séquence d'entrée en terme des unités (mots)

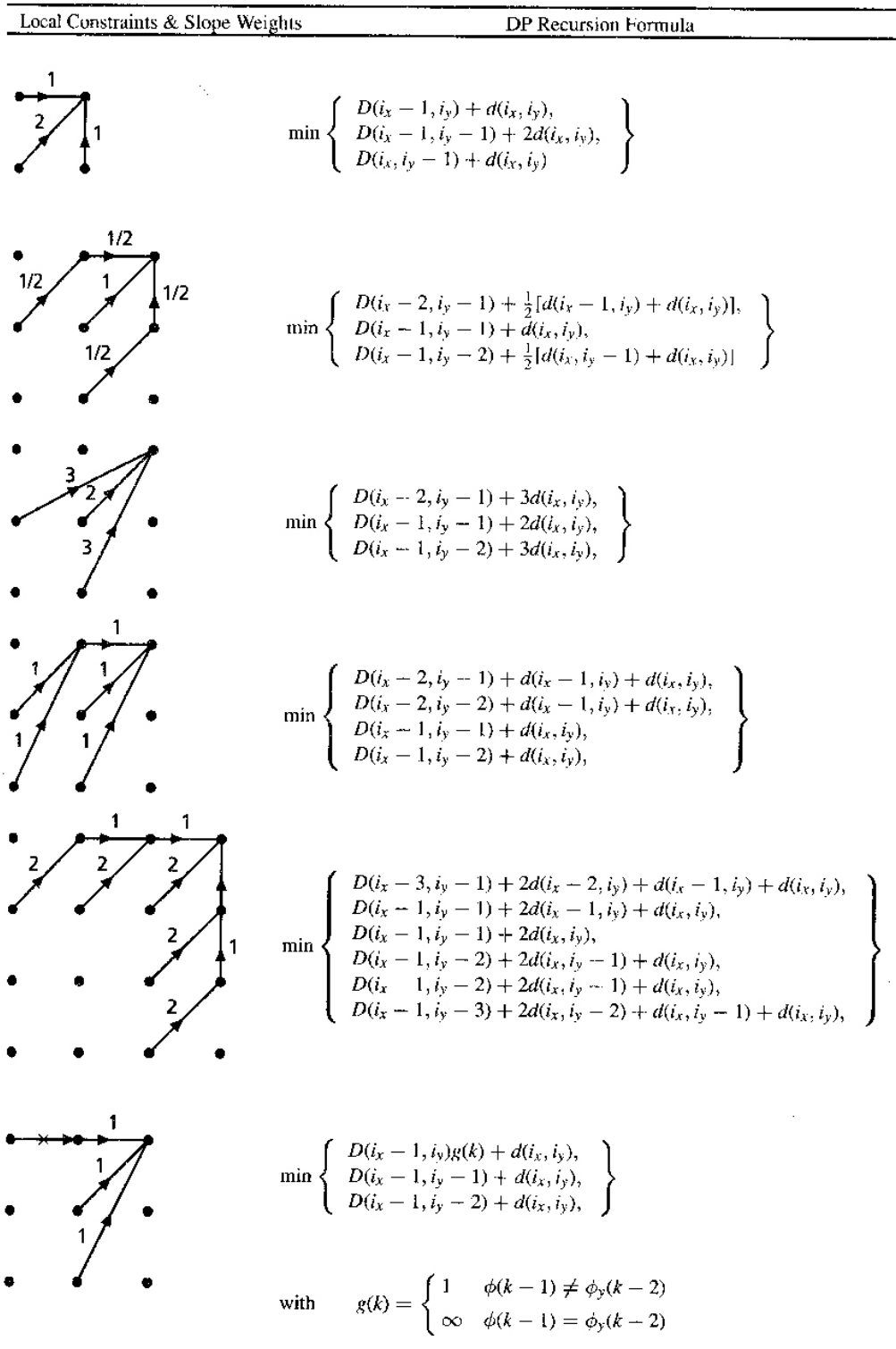


FIGURE III.13 – Contraintes locales et pondération (d'après [22]).

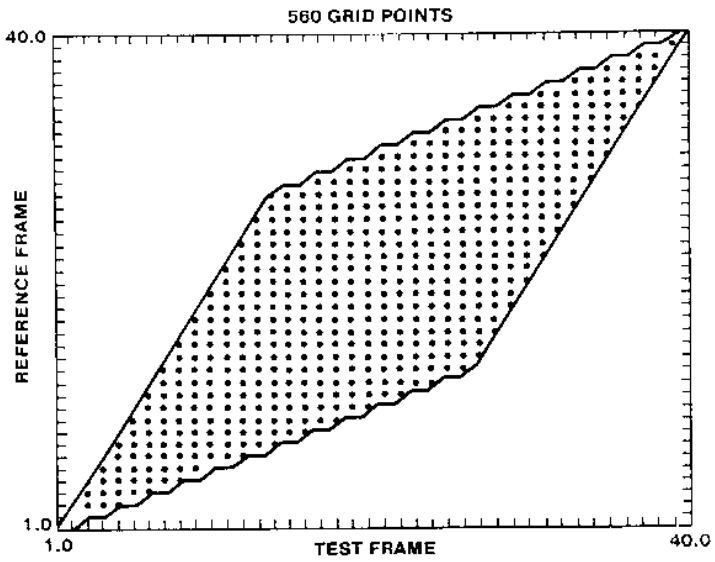


FIGURE III.14 – Région de recherche (contraintes temporelles permettant un taux de compression/expansion local de 2 :1)

de référence. Plusieurs approches ont été proposées pour adapter l'algorithme de programmation dynamique (voir [5],[22]).

Nous ne décrirons ici que l'un d'entre elles, l'approche de programmation dynamique en une passe (*one-pass dynamic time warping*) en raison de sa faible complexité mais aussi parce que c'est l'approche communément adoptée et qu'elle est à la base du décodage de Viterbi utilisé dans les systèmes HMM.

L'algorithme en une passe est très semblable à l'algorithme DTW pour les mots isolés. Cet algorithme, comme pour la reconnaissance de mots isolés, commence par construire une grande matrice de distances locales entre tous les vecteurs constituant les mots de références (les mots du vocabulaire) et tous les vecteurs de la phrase test. On fait alors la programmation dynamique à travers toute la matrice, de gauche à droite, avec les conditions suivantes :

- au départ, le chemin peut commencer à partir de n'importe quel début de mot (en d'autres termes, le chemin ne commence pas nécessairement au point $(1, 1, 1)$ correspondant au point $(1, 1)$ pour le mot de référence 1, mais peut commencer à l'un des points correspondant au début d'une référence soit $(1, 1, k)$ où k représente la k^{ieme} référence)
- à chaque instant n, l'ensemble des successeurs possibles associés au début de chaque mot $(n, 1, k)$ contient également la coordonnée $(n-1, J(k'), k')$ correspondant au dernier indice de tous les mots k' pouvant précéder k.
- à l'intérieur des références, les prédécesseurs possibles sont identiques au cas des mots isolés et dépendent des contraintes locales retenues.

La figure III.15 donne un exemple de chemin DTW dans le cas de mots enchaînés.

III.5.3 Discussion

La programmation dynamique a été utilisée dès les années 1970. C'est cependant dans les années 1980 qu'elle est devenu un standard pour la reconnaissance vocale. L'intégration de distances locales dans le temps est devenue une notion essentielle qui est à la base de tous les systèmes modernes de reconnaissance et notamment ceux basés sur les modèles de Markov cachés.

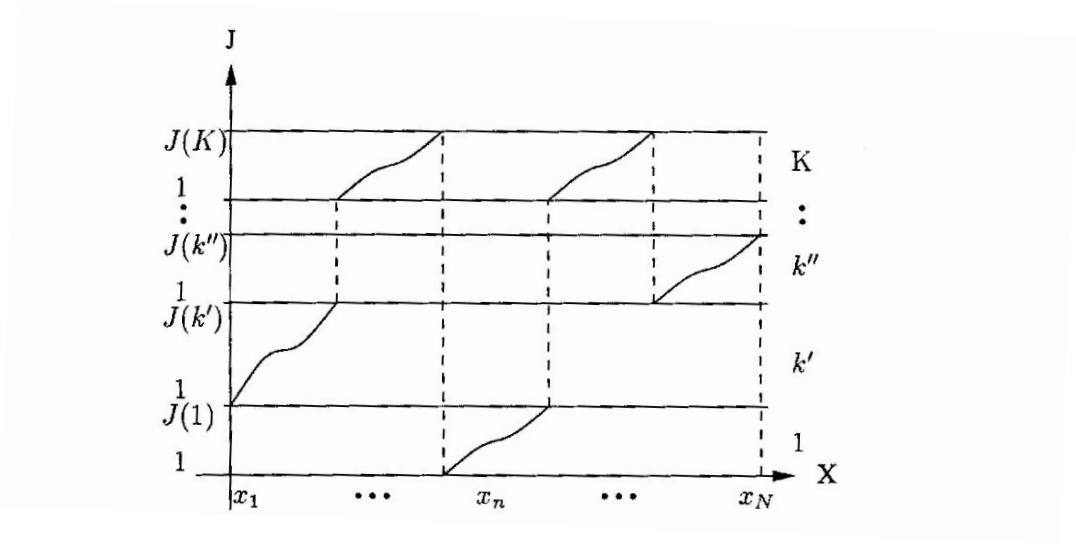


FIGURE III.15 – Exemple de chemin DTW dans le cas de mots enchaînés. Dans cet exemple la phrase prononcée contenait la séquence de mots k' - K - 1 - K - k" (d'après [5]).

De nombreuses variantes et améliorations ont été apportées à ces approches. Notons, que nous avons toujours supposé que chaque mot de vocabulaire n'était représenté que par une seule prononciation. Il est clair qu'en utilisant plusieurs prononciations du même mot permet d'envisager de meilleurs taux puisqu'une certaine variabilité sera alors prise en compte. La solution la plus simple avec l'approche par DTW est de prendre plusieurs références par mot à reconnaitre et d'effectuer plusieurs reconnaissance DTW. Cette solution peut être suffisante pour des systèmes mono-locuteurs mais est vite impraticable pour des systèmes multilocuteurs. L'une des améliorations consiste à utiliser la quantification vectorielle permettant de regrouper soit plusieurs références d'un mot en une seule soit de regrouper les vecteurs acoustiques représentant ces références. On peut, par exemple, utiliser l'algorithme des K-means pour définir des vecteurs de mots prototypes à partir de l'ensemble des vecteurs acoustiques des mots de référence ([6]). Notons qu'il n'est pas ici nécessaire de savoir à quel mot appartiennent les vecteurs acoustiques. Les vecteurs acoustiques acoustiques constituant les mots de référence sont ensuite remplacés par l'étiquette du vecteur prototype le plus proche. Cette quantification vectorielle engendre un certain lissage des références et représente un pas vers les modèles HMM ([5]). Les améliorations majeures apportées à cette approche de base DTW concernent principalement les notions de distances statistiques et les procédures d'entraînements qui y sont liées.

Bibliographie

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

[2] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, New-Jersey, USA., 1957.

[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] H. Bourlard, H. Ney, and C. Wellekens. Connected digit recognition using vector quantization. *In Proc. of ICASSP*, pages 16.10.1–4, 1984.

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

[7] 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.

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

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

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

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

[12] S. Furui. *Digital Speech Processing, Synthesis and Recognition.* Signal Processing and Communictions Series. Marcel Dekker, Inc., 2nd edition edition, 2001.

[13] W. Hartmann. *Signals, Sound and Sensation.* AIP Press, Woodbury, New York, 1997.

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

[15] 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.

[16] 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.

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

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