UE SI340 - Traitement de la Parole -

Gaël RICHARD

11 mai 2010

Table des matières

1. Chapitre reprenant de larges extraits du polycopié de cours de T. Dutoit [27] et du cours de F. Beaugendre [10]

Chapitre I

Introduction

I.1 Introduction

Le domaine des télécommunications s'est considérablement transformé au cours des vingt dernières années et a accompagné la naissance de ce que l'on nomme aujourd'hui *la Société de l'Information ou de la Connaissance* [1]. D'un modèle analogique centré sur le transport de la voix, nous nous dirigeons clairement vers des systèmes de télécommunications numériques centrés sur le transport des données (images, textes, vidéos, documents, etc.) et intégrant des services aux utilisateurs de plus en plus développés. L'explosion d'Internet a été l'un des facteurs clés pour cette évolution et a indubitablement accéléré la convergence du domaine des télécommunications avec celui de l'informatique. Et pourtant pour beaucoup, les principales évolutions sont encore devant nous et seront notamment amenées par la mobilité dans les télécommunications. Les systèmes actuels de téléphonie mobile, dits de seconde génération, sont principalement basés sur le transport de la voix. Les évolutions de ces systèmes autoriseront progressivement le transport des données à des débits suffisants pour faire vivre le concept de l'Internet mobile. Ces évolutions préfigureront ce que seront les systèmes de 3ième génération (tel que l'UMTS) qui permettront des communications mobiles totalement multimédia pour lequel d'innombrables services et applications peuvent être envisagées [[47]].

Ces services peuvent être de tout ordre et incluent notamment les services de renseignement (navigation automobile, météo, annuaire,. . .), les services de messagerie ou d'accueil (répondeurs, call centers, messagerie unifiée,. . .) et les services financiers (banque, bourse, commerce électronique,. . .). Le succès de ces services automatisés dépendra en grande partie de facteurs clés tels que :

- *L'accessibilité:* c'est à dire la capacité d'accéder au service de n'importe quel endroit et en particulier en situation de mobilité.
- *L'interopérabilité:* c'est-à-dire la possibilité d'accéder au service à partir de tout type de terminal (agenda électronique, téléphone portable, ordinateur,)
- *La facilité d'utilisation* ou plus précisément la facilité avec laquelle un utilisateur n'ayant pas de connaissances particulières pourra accéder au service.
- *La confidentialité* ou sécurité des accès, notamment pour les services financiers.

La facilité d'utilisation doit se traduire par un accès intuitif ou encore mieux un accès naturel au service, c'est à dire à travers des interfaces intégrant les modes de communications naturels que l'homme utilise pour communiquer ou dialoguer. C'est là l'enjeu des interfaces Homme-Machine qui peuvent proposer de nouveaux moyens d'accès naturels et intuitifs à l'information, malgré la complexité des technologies sous-jacentes.

L'étendue des recherches visant à améliorer les interfaces Homme-Machine est très grande et inclue les différentes modalités de communication: parole, écrit, visuel, graphique, gestuel ainsi que leur combinaison dans des systèmes multimodaux ou dans des systèmes de réalité virtuelle ou augmentée (voir par exemple [21, 15, 84, 33, 17]).

Dans ce cadre très général, il est indubitable que la parole tient une place très importante puisqu'elle est le moyen de communication privilégié de l'homme mais qu'elle représente aussi un vecteur primordial pour la miniaturisation (accès directe à l'information sans clavier).

La parole, étant centrale dans la communication, a toujours reçu une attention particulière des hommes pour mieux la comprendre, mieux la représenter mais aussi mieux la transporter. Cependant, la parole est un objet complexe et ne peut être décrit ou analyser à travers une vision unique. Elle est en effet par nature à la croisée de multiples disciplines incluant la physique, l'acoustique, la linguistique, la perception, la médecine, et le traitement du signal.

Comprendre les phénomènes mis en jeu durant la production et la perception de la parole est essentiel pour son traitement automatique et c'est pour cela qu'une partie relativement importante leur sont consacrés.

On distingue usuellement 3 grandes familles d'application du traitement de la parole:

- *Le codage:* qui vise à compresser le signal de parole en vue de sa transmission ou de son stockage.
- *La synthèse vocale:* qui vise à produire un signal de parole à partir d'une écriture symbolique (texte)
- *La reconnaissance vocale:* qui vise à transcrire le signal vocal en texte.

Dans ce cours, nous aborderons les aspects liés à la synthèse et à la reconnaissance de la parole mais nous ne traiterons pas du codage de la parole (le lecteur intéressé par ce domaine passionnant pourra, par exemple, consulter [70]). Ce document est ainsi organisé comme suit. La première partie de ce document (Chapitre II) sera consacrée à la description des phénomènes de Production et de Perception de la parole. On abordera dans un second temps (Chapitre III), la modélisation acoustico-articulatoire qui représente une première direction pour la synthèse de parole. La partie suivante (Chapitre IV) sera dédiée à la reconnaissance vocale et présentera les principaux concepts de ce domaine. On abordera enfin (Chapitre V) les différents aspects de la synthèse de parole à partir du texte.

Chapitre II

Production et Perception de la parole

II.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.

II.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 II.1). La respiration est un phénomène mécanique intégrant une phase active (l'inspiration) et une phase passive (l'expiration).

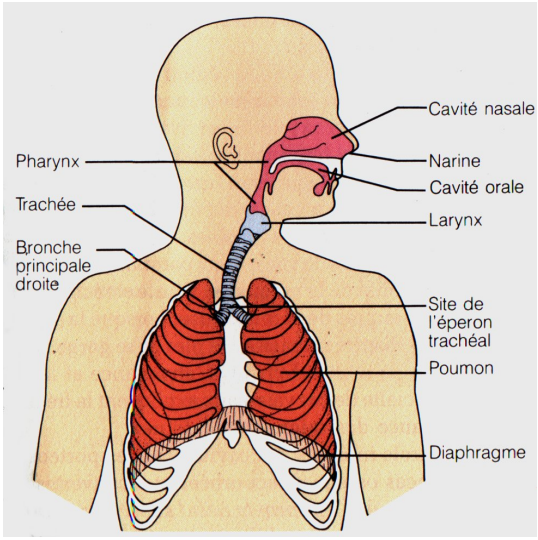


FIG. II.1 – *Schéma de l'appareil respiratoire (d'après [23])*

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 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 [14].

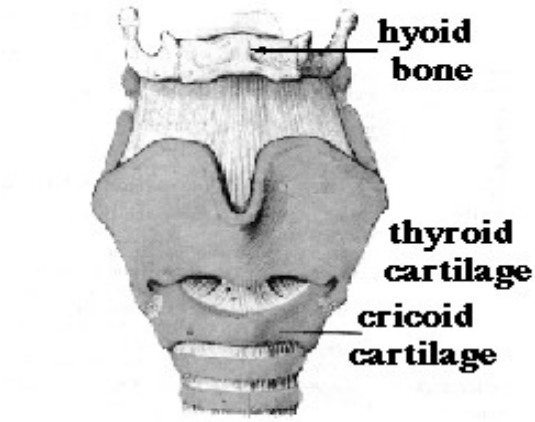


FIG. II.2 – *Schéma du larynx (d'après [75])*

II.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 II.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

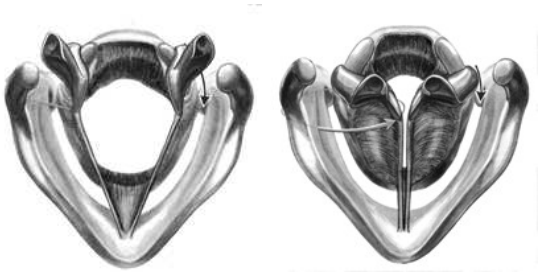


FIG. II.3 – *Les cordes vocales en position ouvertes durant la respiration (à gauche) et fermées pour la production de parole (à droite),(d'après [75])*

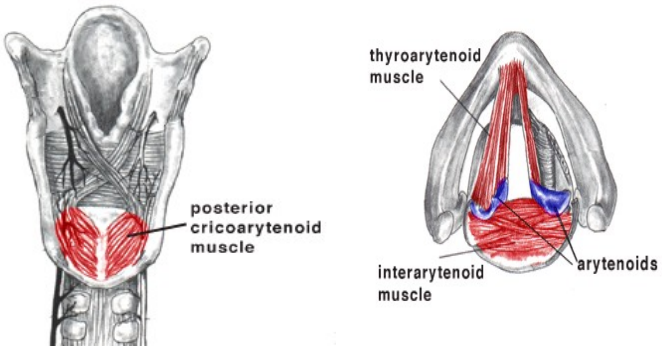


FIG. II.4 – *Schéma des muscles intrinsèques du larynx (d'après [75])*

plutôt l'espace se trouvant entre les cordes vocales). Les cordes vocales sont particulièrement 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 II.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 II.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.

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 ,

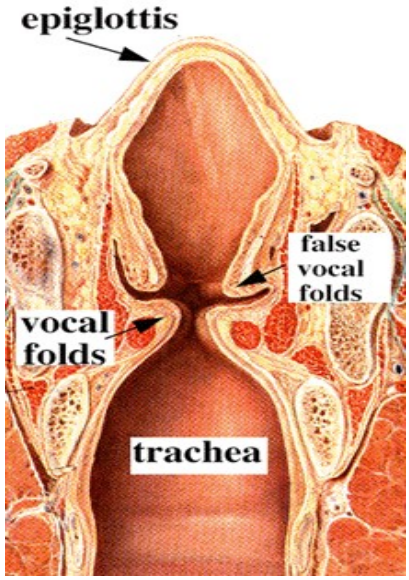


FIG. II.5 – *Vue longitudinale du larynx (d'après [75])*

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 II.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 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.

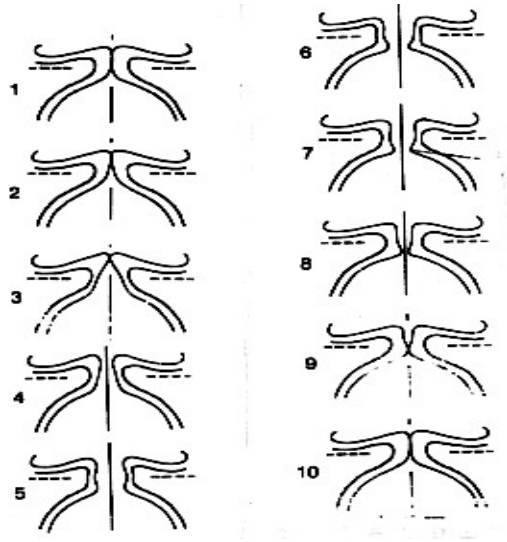


FIG. II.6 – *Schéma de vibration des cordes vocales (d'après [75])*

Le diagramme ci-dessus (figure II.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 [32] ou [18].

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.

1. Notons que le terme bruit d'aspiration est parfois réservé au bruit émis au niveau (ou prés) de la glotte ([92], [31]).

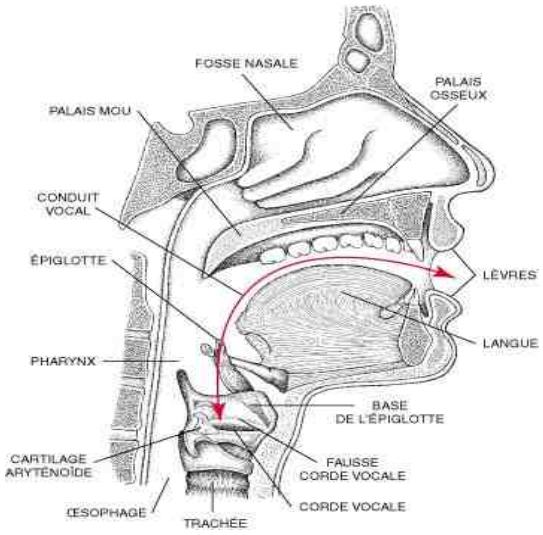


FIG. II.7 – *Vue schématique du conduit vocal humain d'après [57]*

– 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.

II.1.3 Les cavités supraglottiques

Il existe 2 cavités supraglottiques (v. figure II.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 II.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 [14].

La langue mobile a la forme d'une pyramide à faces arrondies, constituée dune 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 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 [14] 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 intéro-labial 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

II.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 [18] 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 [49]) dont a été extrait le tableau complet de l'alphabet phonétique international donné figure II.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.

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 II.9 donne une classification des phonèmes du français suivant ces traits distinctifs généraux.

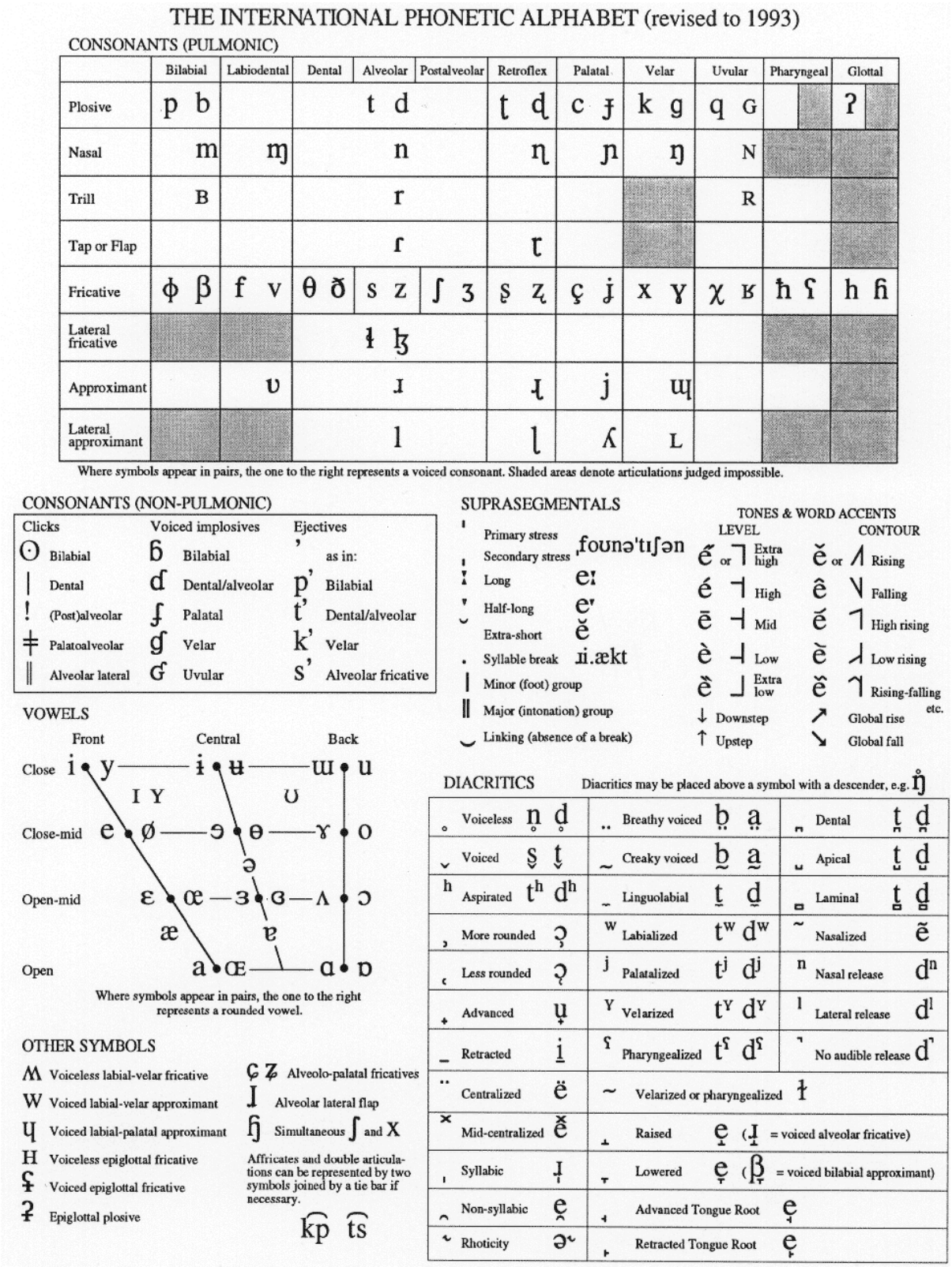


FIG. II.8 – *Tableau complet de l'alphabet phonétique international [49]*

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 Non arrondies		Postérieures Arrondies
Fermées	[i]	[y]	[u]
	[e]	[ø]	[o]
	[ɛ]	[œ]	[ɔ]
Ouvertes	[a]		
Nasales	Antérieures		Postérieures
Fermées	[ɛ̃]		[õ]
Ouvertes		[ã]	

FIG. II.9 – *Classification des phonèmes du français [18]*

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:

- 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 :

- 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/.

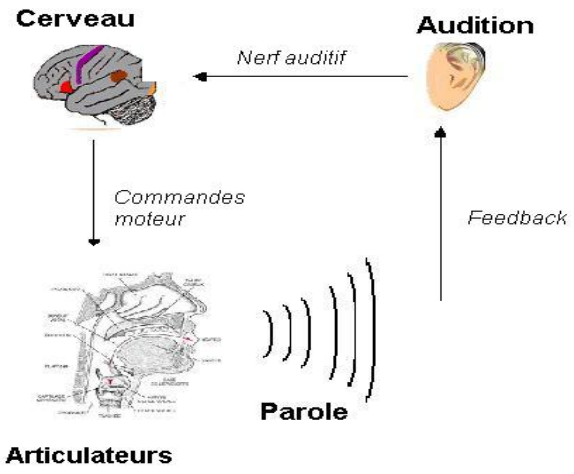


FIG. II.10 – *Système de production et feedback auditif*

- 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".

II.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 II.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.

II.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). ([22])

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).

II.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 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 II.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.

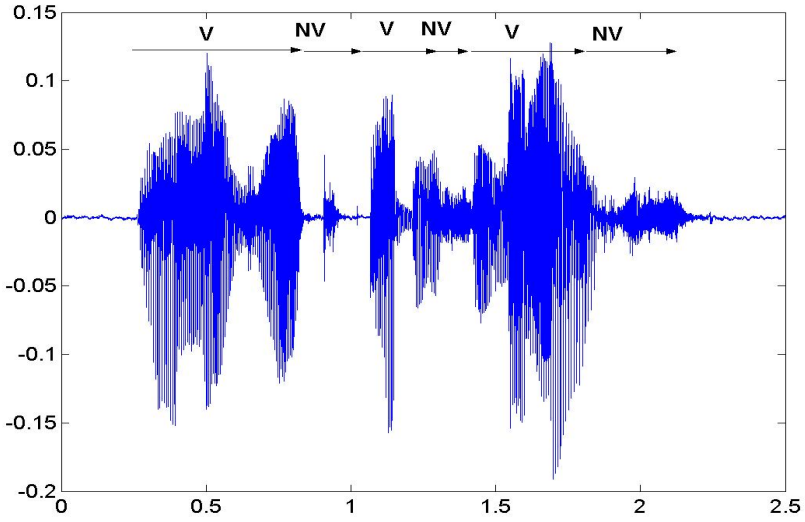


FIG. II.11 – *Signal temporel de la phrase "La musique adoucit les moeurs": (V=partie voisée; NV = partie non voisée*

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]:

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

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

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

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

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

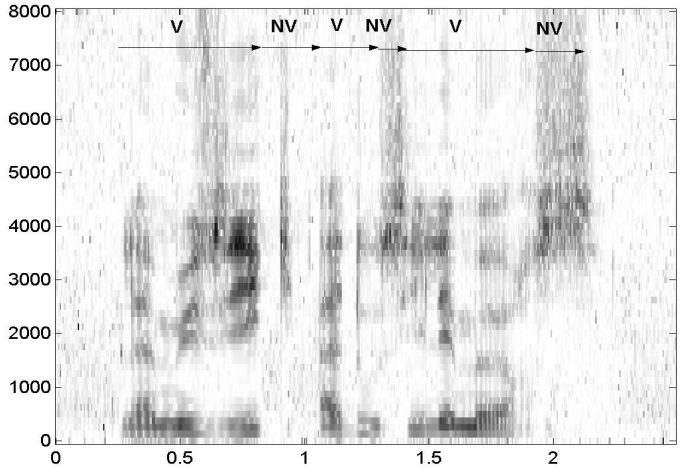


FIG. II.12 – *Spectrogramme de la phrase "La musique adoucit les moeurs" (Le spectrogramme représente le module de la transformée de Fourier où cours du temps avec les Fréquences en ordonnée, le temps en abscisse et l'énergie en niveau de gris. Ainsi une zone sombre, indique une forte énergie à la fréquence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

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 II.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 II.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 II.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.

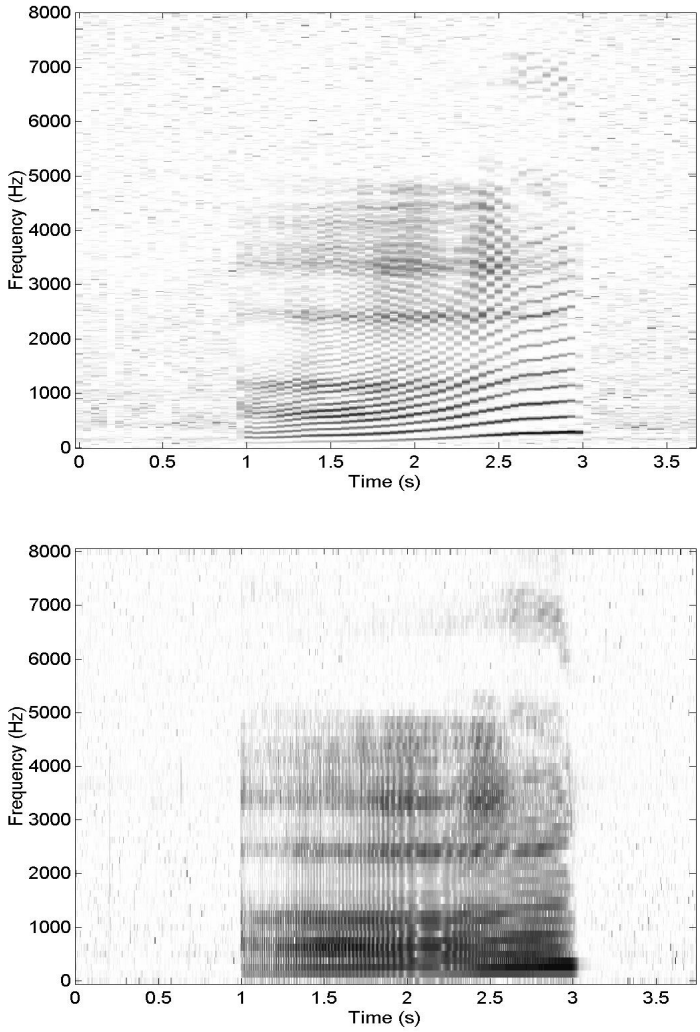


FIG. II.13 – *Spectrogramme bande étroite (haut) et spectrogramme large bande (bas) d'une voyelle /a/ produite avec une élevation progressive de la fréquence fondamentale"*

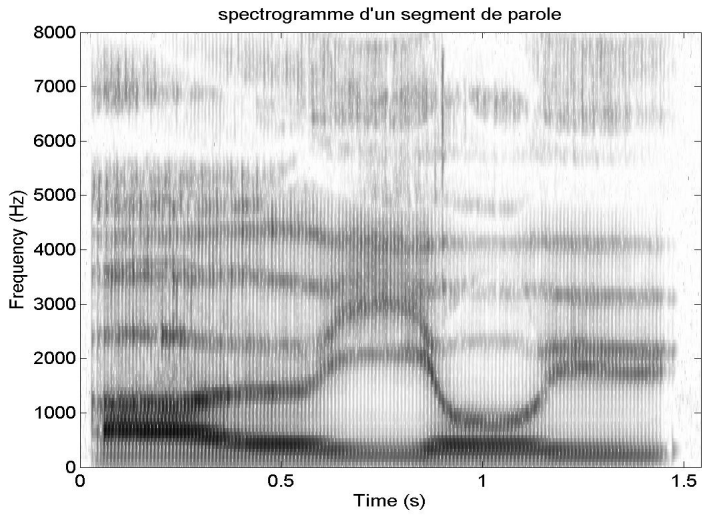


FIG. II.14 – *Spectrogramme du son constitué des voyelles /aeiou/: les mouvements brusques des deux premiers formants, indiquent un changement de voyelle*

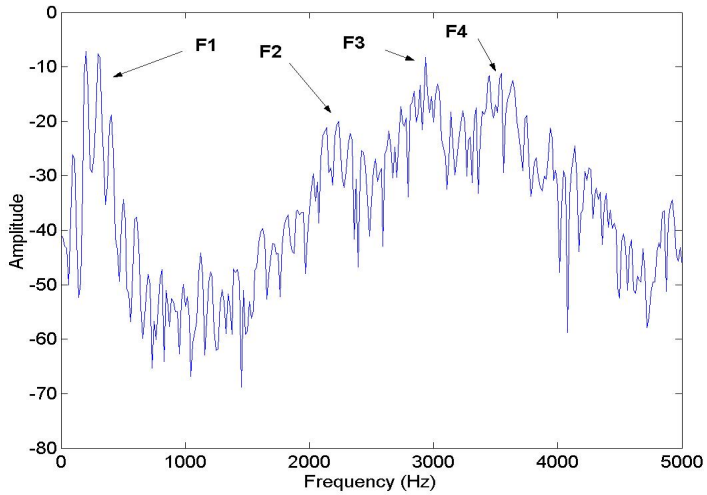


FIG. II.15 – *Module de la transformée de Fourrier (ou coupe spectrographique) d'une trame de la voyelle /i/*

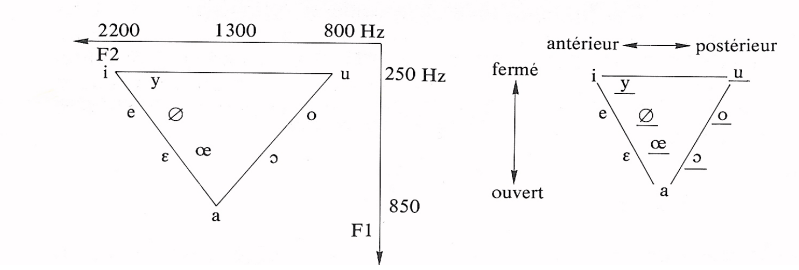
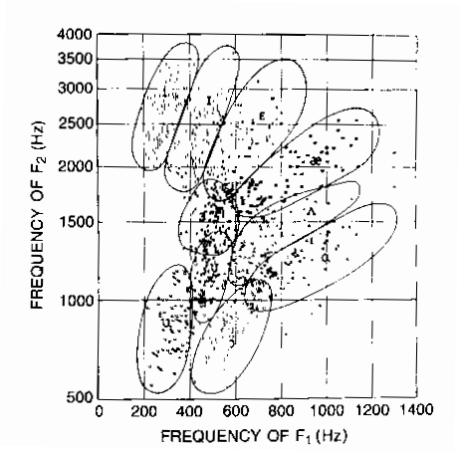


FIG. II.16 – *Triangle vocalique*



In *Fundamentals of speech recognition*, L. Rabiner & B-H. Juang, ©Prentice Hall, 1993

FIG. II.17 – *Représentation des sons vocaliques de l'anglais en fonctions des deux premières fréquences formantiques (D'après [80])*

près du palais donnant lieu à une constriction plus étroite , voir figure II.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 II.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.

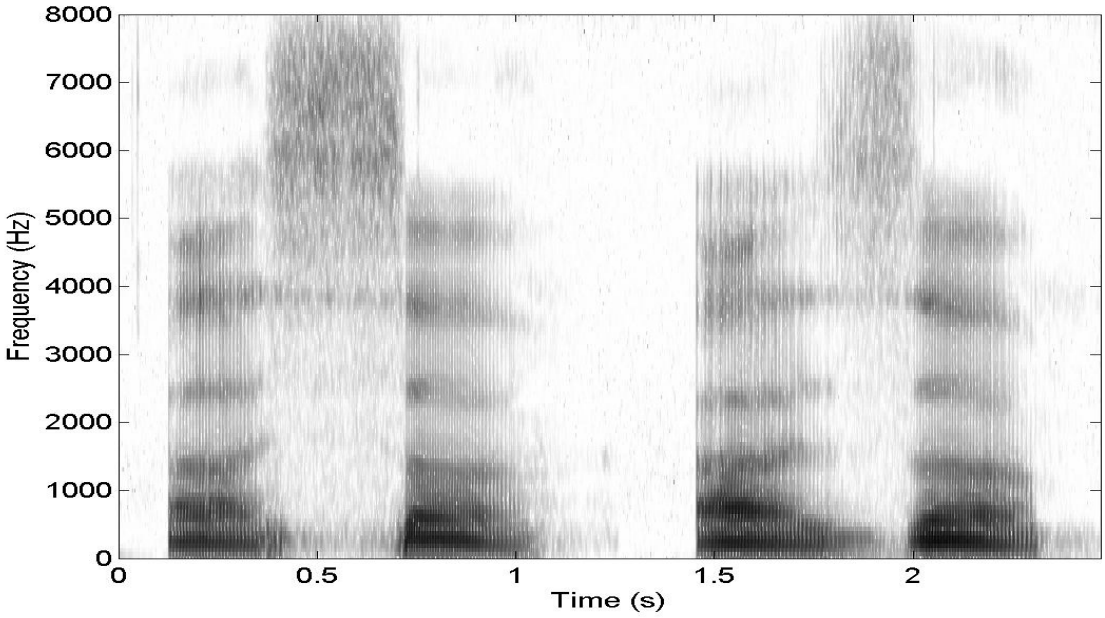


FIG. II.18 – *Spectrogramme bande large du signal "assa aza" (/*a s a a z a*/)"*

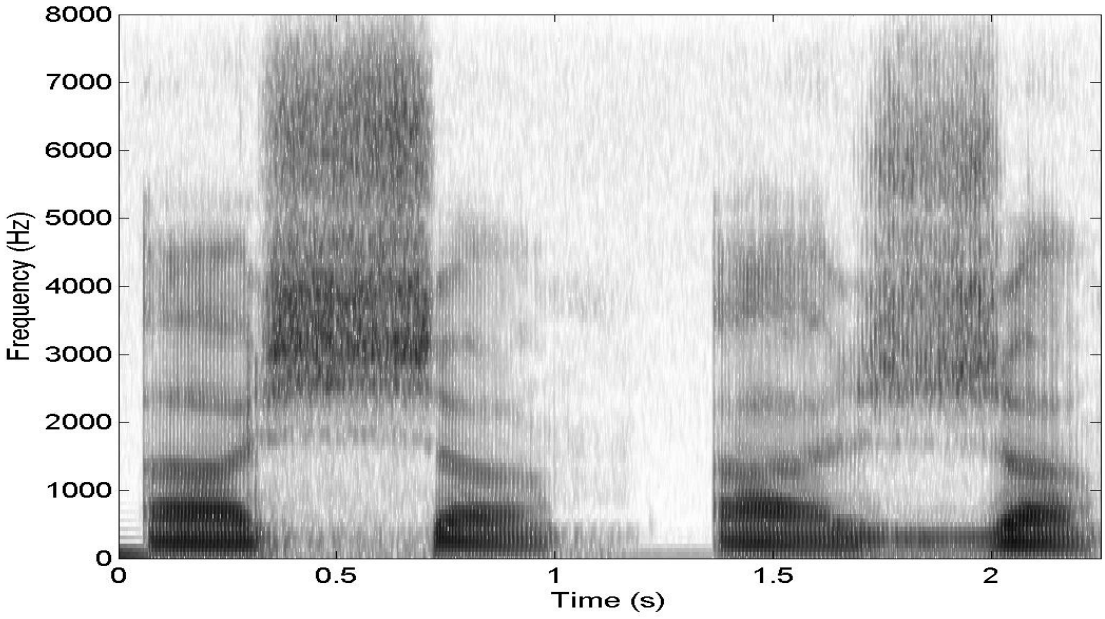


FIG. II.19 – *Spectrogramme bande large du signal "acha aja" (/*a ʃa a ʒa*/)"*

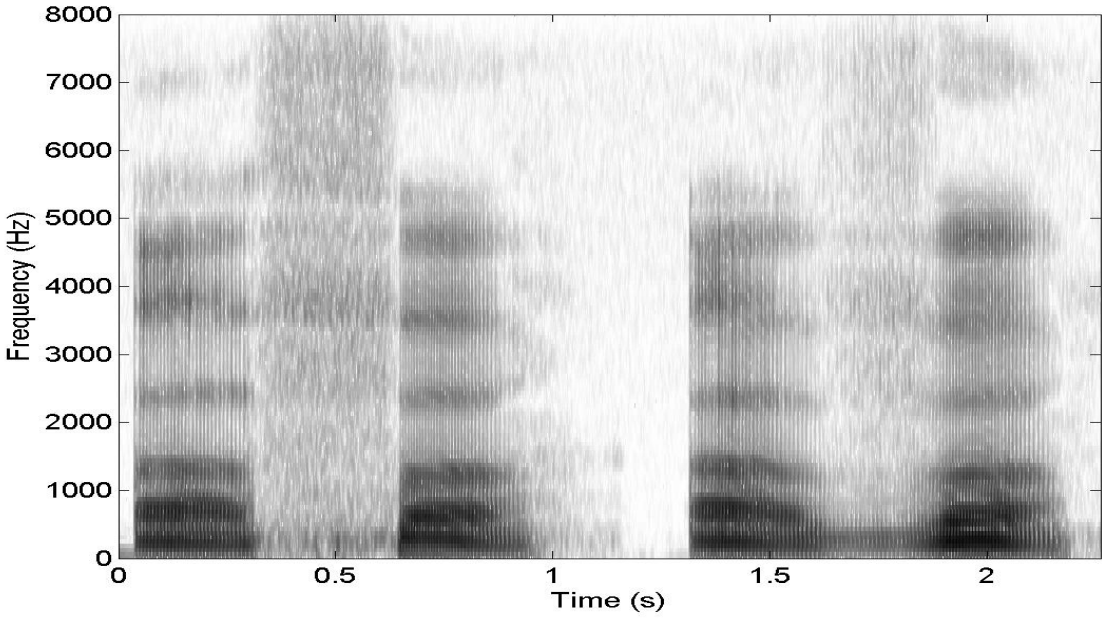


FIG. II.20 – *Spectrogramme bande large du signal "afa ava" (/*a f a a v a*/)"*

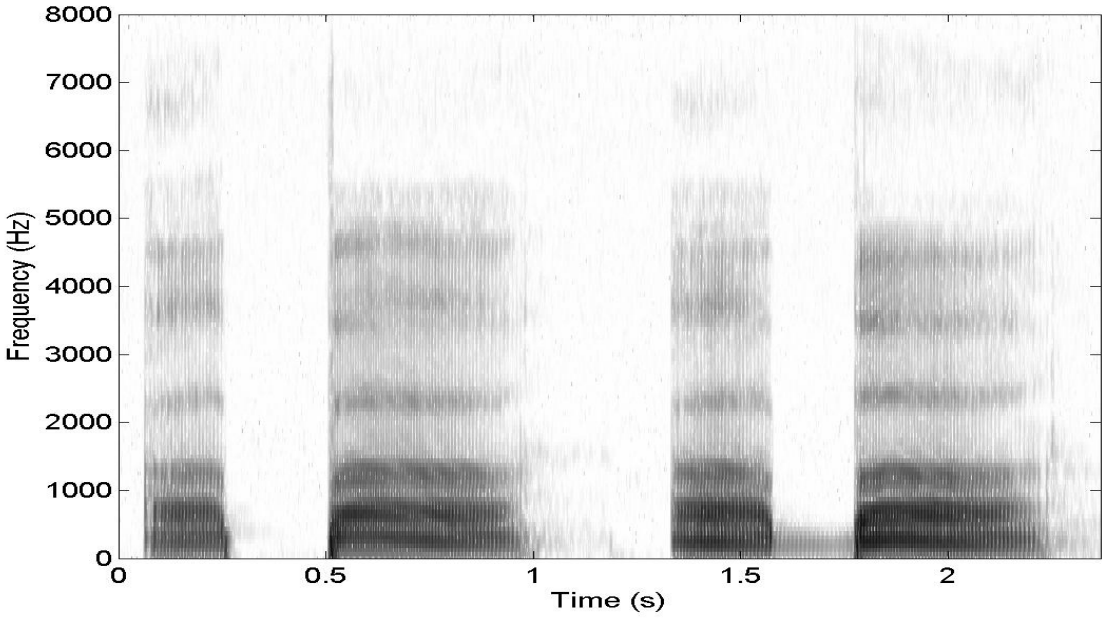


FIG. II.21 – *Spectrogramme bande large du signal "apa aba" (/*a p a a b a*/)"*

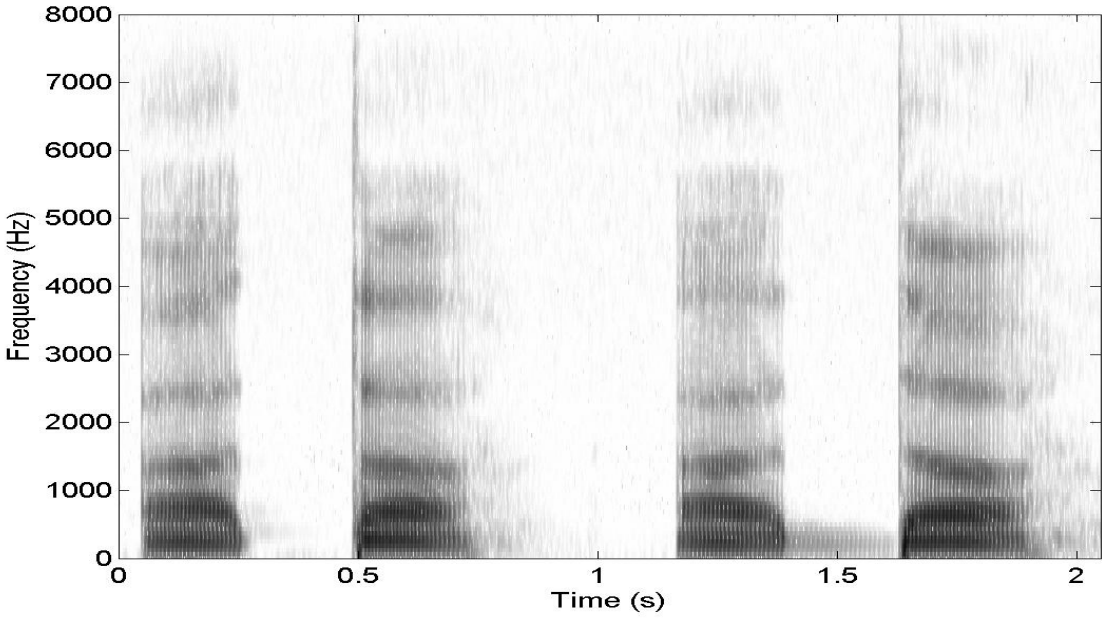


FIG. II.22 – *Spectrogramme bande large du signal "ata ada" (/*a t a a d a*/)"*

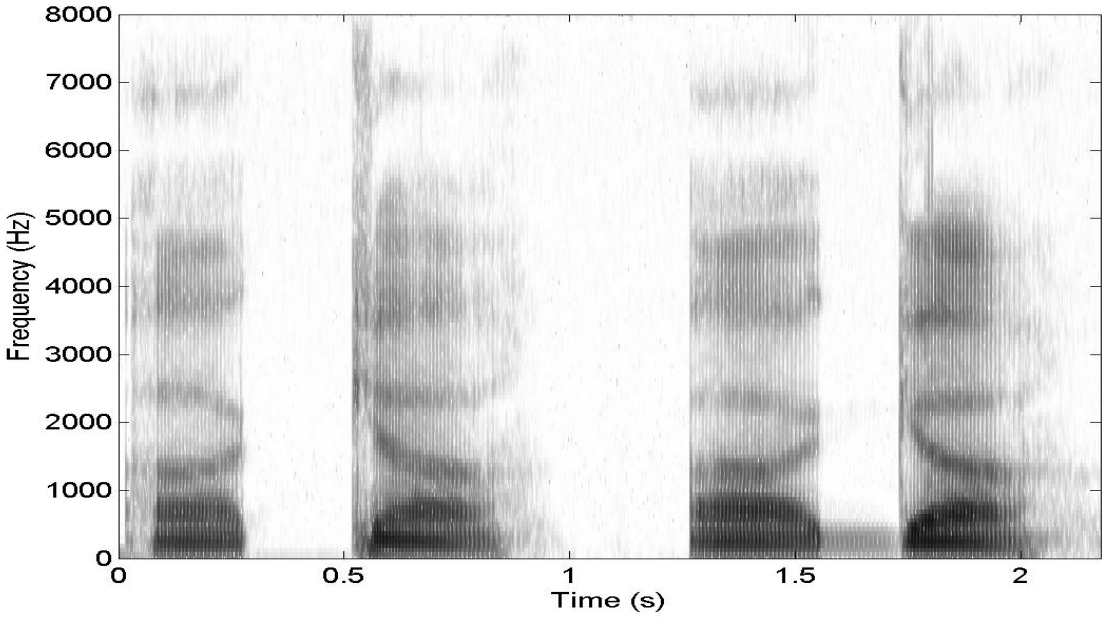


FIG. II.23 – *Spectrogramme bande large du signal "aka aga" (/*a k a a g a*/)"*

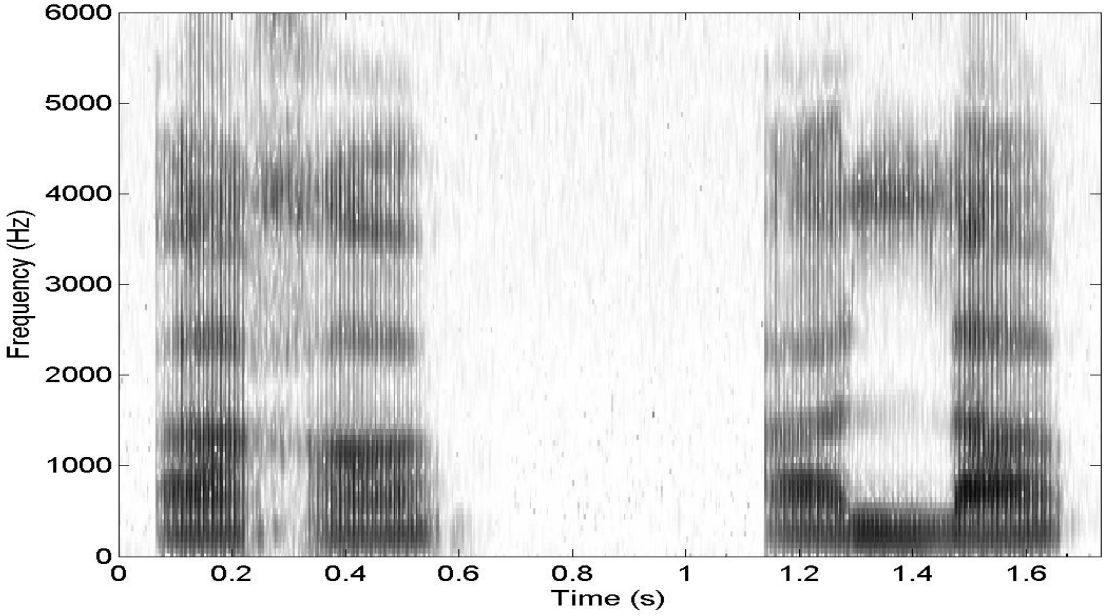


FIG. II.24 – *Spectrogramme bande large du signal "ara ala" (/*a R a a l a*/)"*

Chapitre III

Modélisation articulatoire

Nous avons vu au cours des chapitres précédents des notions de production et de perception de la parole et l'utilité de la représentation spectrale pour décrire les caractéristiques acoustiques de ces sons.

La modélisation articulatoire vise essentiellement à modéliser le processus de production de la parole. Cette modélisation est relativement complexe et comporte de nombreuses facettes allant d'une modélisation précise d'un articulateur ou d'une source sonore jusqu'à des approches plus globales visant plutôt des objectifs de synthèse vocale.

La modélisation articulatoire est relativement ancienne et un grand nombre de modèles et d'approches ont été développées. L'objectif de ce cours n'est pas de présenter ces méthodes de façon exhaustives mais plutôt de présenter quelques modèles et de renvoyer le lecteur à la littérature du domaine pour approfondir ses connaissances.

L'un des concepts les plus importants décrivant la structure du signal de parole est déduit du modèle source-filtre de la parole ([31]). Dans ce modèle le signal de parole est vu comme une (ou plusieurs) source(s) sonore(s) qui a (ont) subi des modifications spectrales selon la forme du conduit vocal qui agit comme un filtre acoustique. Comme nous l'avons vu, une source vocale peut-être soit voisée (lorsqu'il y a vibration des cordes vocales) soit non voisée lorsqu'un bruit turbulent est créé au niveau d'une constriction du conduit vocal ou suite à un relâchement d'une occlusion.

III.1 Théorie acoustique

On peut considérer le conduit vocal comme un tube acoustique. La forme de ce tube est géométriquement complexe. Les techniques modernes d'imagerie médicales telles que les Images à Résonance Magnétique (IRM) permettent d'obtenir des données physiologique dynamiques (c'est à dire pendant l'élocution) de très bonne qualité.

Nous donnons figure III.1 un exemple d'images IRM lors de la production d'un /i/ et d'un /f/ [29]. A partir de ces images, il est possible de reconstruire la forme tridimensionnelle du conduit vocal (voir figure III.2 pour un exemple de reconstruction pour le son /f/)

Le conduit vocal forme un angle de 90 degrés en son milieu. Cette courbure n'a cependant pas d'effet acoustique pour des fréquences inférieures à 5 kHz. Ainsi, pour la modélisation nous pourrons "redresser" le conduit vocal, et le représenter comme un tube droit sans courbure. Par ailleurs, le mode principal de propagation des ondes sonores à l'intérieur d'un tube est longitudinal (cette hypothèse est vérifiée principalement pour les fréquences inférieures à 4 kHz. Nous pourrons ainsi ignorer le mode de résonance transversal et supposer que nous avons une

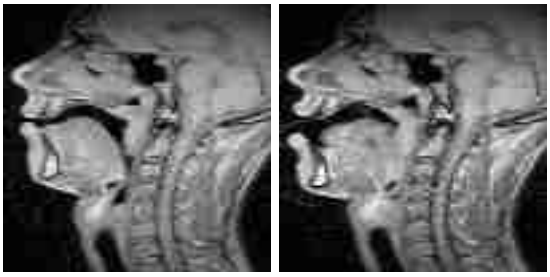


FIG. III.1 – *Images IRM du conduit vocal pour la fricative /f/ (à gauche) et la voyelle /a/ (à droite) suédoises (d'après The 3D vocal tract project [29])*

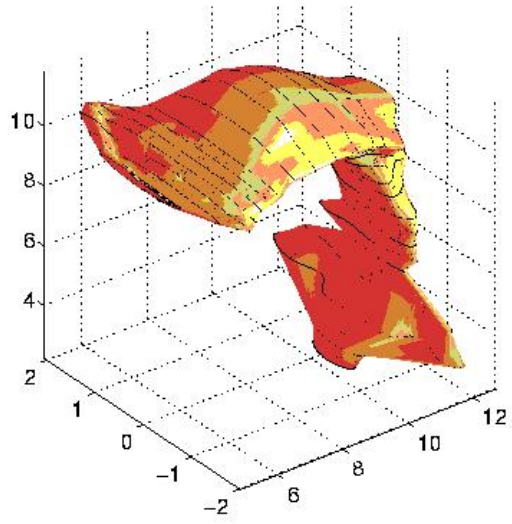


FIG. III.2 – *Reconstruction de la forme tridimensionnelle du conduit vocal durant la phonation du son /fi/ à partir d'images IRM (d'après [29])"The 3D vocal tract project")*

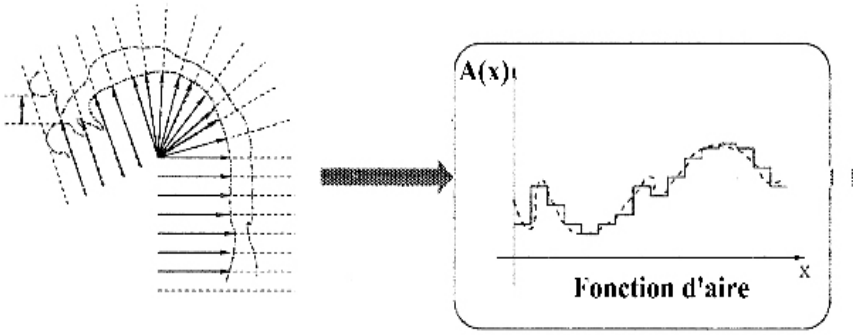


FIG. III.3 – *Fonction d'aire déduite à partir d'un modèle articulatoire utilisant une grille semi-polaire*

propagation en ondes planes.

Finalement, en supposant que la section transversale est circulaire, nous pouvons grâce aux hypothèses ci-dessus en déduire un tube acoustique équivalent qui sera droit et de section circulaire. En raison de la symétrie de ce tube, on représentera souvent le conduit vocal par sa fonction d'aire qui spécifie la variation de la section transversale de la glotte aux lèvres.

On donne figure III.3 un exemple de fonction d'aire obtenu à partir d'un modèle articulatoire (modèle de S. Maeda [62]).

III.1.1 Les équations fondamentales

Soit, $A(x,t)$ cette fonction d'aire variant dans le temps où x est la distance à partir de la glotte et t le temps.

Le champ acoustique est définit en tout point x et à chaque instant t par :

– la pression: $P(x,t)$

– la vitesse de l'air : $U(x,t)$

Considérons un élément d'air unitaire de masse m et de volume V. Sa densité volumique ρ s'écrit:

$$\rho = \frac{m}{V} \tag{III.1}$$

En supposant qu'il règne en ce point la pression $P(x,t)$, l'élément considéré sera soumis à la force :

$$F = -\frac{\partial P}{\partial x} V \tag{III.2}$$

L'équation de mouvement

On peut alors écrire pour cet élément, l'équation du mouvement (ou seconde loi de Newton) qui est parfois appelée loi de conservation de mouvement:

$$F = m\frac{dU}{dt} \tag{III.3}$$

Nous pouvons exprimer dU en fonction de ses dérivées partielles:

$$dU = \frac{\partial U}{\partial x}dx + \frac{\partial U}{\partial t}dt \tag{III.4}$$

En utilisant les équations III.2, III.4 et III.1, on obtient l'équation du mouvement suivante:

$$-\frac{\partial P}{\partial x} = \rho\frac{\partial U}{\partial t} + \rho U\frac{\partial U}{\partial x} \tag{III.5}$$

L'équation de conservation de la masse

De même, on peut écrire pour l'élément d'air unitaire, l'équation de conservation de la masse (parfois appelée équation de continuité) qui traduit qu'au cours du passage de la section $A(x)$ à la section $A(x+dx)$ la masse m, délimitée dans le volume V se conserve totalement. Cette équation est obtenue :

- en écrivant la variation de masse au passage de $A(x)$ à $A(x+dx)$: $-\frac{\partial}{\partial x}\int\int_{A(x)} \rho U dA$
- en écrivant la masse de l'élément qui s'est déplacé entre x et $x+dx$ pendant dt, soit $\frac{\partial}{\partial t}\int\int_{A(x)} \rho dA$
- puis en égalant ces deux expressions qui correspondent à la même masse:

$$-\frac{\partial}{\partial x}\int\int_{A(x)} \rho U dA = \frac{\partial}{\partial t}\int\int_{A(x)} \rho dA \tag{III.6}$$

Les équations de Webster

Nous allons maintenant simplifier les équations III.6 et III.5 en faisant un certain nombre d'hypothèses. Tout d'abord, nous allons supposer que l'air est un gaz parfait et que de plus nous sommes en présence d'une transformation adiabatique. Ceci nous permet d'écrire que :

$$PV^\gamma = Cste \tag{III.7}$$

et ainsi de relier la masse volumique ρ au gradient de pression par rapport à la pression atmosphérique notée p (voir poly PAMU-ACOUS):

$$\rho = \rho_0(1 + \frac{p}{\rho_0 c^2}) \tag{III.8}$$

Ainsi, le terme de droite de l'équation III.6 s'écrit:

$$\frac{\partial}{\partial t}\int\int_{A(x)} \rho dA = \frac{\partial}{\partial t}\int\int_{A(x)} \rho_0(1 + \frac{p}{\rho_0 c^2})dA \tag{III.9}$$

$$= \rho_0(\frac{\partial A}{\partial t} + \frac{1}{\rho_0 c^2}\frac{\partial}{\partial t}\int\int_{A(x)} p dA) \tag{III.10}$$

En faisant l'hypothèse supplémentaire que nous sommes en présence d'ondes planes, c'est à dire que la pression p est constante sur une tranche de section A, c'est à dire que nous avons $pA = \int\int_{A(x)} pdA$, nous obtenons:

$$\frac{\partial}{\partial t}\int\int_{A(x)} \rho dA = \rho_0(\frac{\partial A}{\partial t} + \frac{1}{\rho_0 c^2}\frac{\partial (pA)}{\partial t}) \tag{III.11}$$

De même, il est possible de simplifier le membre de gauche qui s'écrit à l'aide de l'équation III.8:

$$-\frac{\partial}{\partial x}\int\int_{A(x)}\rho U dA = -\rho_0(\frac{\partial}{\partial x}\int\int_{A(x)}(UdA + \frac{U}{\rho_0 c^2}pdA)) \quad \text{(III.12)}$$

En faisant l'hypothèse que la vitesse des particules d'air U s'exprime en fonction d'un flux constant et d'une petite variation de vitesse v autour de ce flux constant, c'est à dire que:

$$U = U_0 + v \quad (v \ll U_0) \quad \text{(III.13)}$$

l'équation III.12 se réécrit alors:

$$-\frac{\partial}{\partial x}\int\int_{A(x)}\rho U dA = -\rho_0(\frac{\partial(U_0 A)}{\partial x} + \frac{\partial}{\partial x}\int\int_{A(x)}(vdA + \frac{U_0}{\rho_0 c^2}pdA)) \quad \text{(III.14)}$$

En supposant de plus que toutes les particules d'air sont en place (qui est également une hypothèse d'ondes planes), c'est à dire que $vA = \int\int_{A(x)} vdA$

Nous obtenons alors pour le terme de gauche de l'équation de continuité:

$$-\frac{\partial}{\partial x}\int\int_{A(x)}\rho U dA = -\rho_0(\frac{\partial(U_0 A + vA)}{\partial x} + \frac{1}{\rho_0 c^2}\frac{\partial(U_0 pA)}{\partial x}) \quad \text{(III.15)}$$

$$= -\rho_0(\frac{\partial(U_0 A + vA)}{\partial x} + \frac{U_0}{\rho_0 c^2}\frac{\partial(pA)}{\partial x}) \quad \text{(III.16)}$$

Finalement, en utilisant les expressions III.15 et III.11, l'équation III.6 se réécrit:

$$-\frac{\partial(U_0 A + vA)}{\partial x} - \frac{U_0}{\rho_0 c^2}\frac{\partial(pA)}{\partial x} = \frac{\partial A}{\partial t} + \frac{1}{\rho_0 c^2}\frac{\partial(pA)}{\partial t} \quad \text{(III.17)}$$

De même, en utilisant les hypothèses faites ci-dessus, l'équation de mouvement peut être simplifiée:

$$-\frac{\partial P}{\partial x} = \rho\frac{\partial(U_0 + v)}{\partial t} + \rho(U_0 + v)\frac{\partial(U_0 + v)}{\partial x} \quad \text{(III.18)}$$

$$= \rho\frac{\partial v}{\partial t} + \rho U_0\frac{\partial v}{\partial x} \quad \text{(III.19)}$$

En supposant de plus, que la pression p est petite devant la pression atmosphérique, c'est à dire que nous avons

$$P = P_0 + p \quad (p \ll P_0) \quad \text{(III.20)}$$

nous obtenons l'équation:

$$-\frac{\partial P}{\partial x} = \rho_0\frac{\partial v}{\partial t} + \rho_0 U_0\frac{\partial v}{\partial x} \quad \text{(III.21)}$$

En supposant que les parois sont rigides (soit $\frac{\partial A}{\partial t} = 0$), et que l'on considère des sections constantes (soit $\frac{\partial A}{\partial x} = 0$) on obtient le système d'équations:

$$-\frac{\partial v}{\partial x} - \frac{U_0}{\rho_0 c^2}\frac{\partial p}{\partial x} = \frac{1}{\rho_0 c^2}\frac{\partial p}{\partial t} \quad \text{(III.22)}$$

$$-\frac{\partial P}{\partial x} = \rho_0\frac{\partial v}{\partial t} + \rho_0 U_0\frac{\partial v}{\partial x} \quad \text{(III.23)}$$

Finalement, il est possible de négliger les termes convectifs $U_0 \frac{\partial}{\partial x}$ lorsque la vitesse du flux d'air moyen U_0 est petite devant la vitesse du son c. En effet, nous savons que pour un tube à parois rigides, d'aire constante par morceaux, le système d'équations ci-dessous admet pour solution une superposition d'ondes dans chaque section sous la forme:

$$p(x,t) = p_+(x-ct) + p_-(x+ct) \quad \text{(III.24)}$$
$$v(x,t) = v_+(x-ct) + v_-(x+ct) \quad \text{(III.25)}$$

Or pour toute fonction $f(x+ct)$, nous avons:

$$\frac{\partial(f(x+ct))}{\partial t} = cf'(x+ct) \quad \text{(III.26)}$$
$$U_0 \frac{\partial(f(x+ct))}{\partial x} = U_0 f'(x+ct) \quad \text{(III.27)}$$

Ainsi, si $U_0 \ll c$ les termes convectifs $U_0 \frac{\partial}{\partial x}$ peuvent être négligés par rapport aux termes en $\frac{\partial}{\partial t}$. On obtient alors les fameuses équations de Webster:

$$-\frac{\partial p}{\partial x} = \rho_0 \frac{\partial v}{\partial t} \quad \text{(III.28)}$$
$$-\frac{\partial v}{\partial x} = \frac{1}{\rho_0 c^2} \frac{\partial p}{\partial t} \quad \text{(III.29)}$$

qui peuvent être écrite en fonction de la vitesse volumique $U_v = \frac{v}{A}$:

$$-\frac{\partial p}{\partial x} = \frac{\rho_0}{A} \frac{\partial U_v}{\partial t} \quad \text{(III.30)}$$
$$-\frac{\partial U_v}{\partial x} = \frac{A}{\rho_0 c^2} \frac{\partial p}{\partial t} \quad \text{(III.31)}$$

Bien entendu, le conduit vocal est bien plus complexe qu'un unique tube de section uniforme. En pratique, on pourra cependant approximer le conduit vocal par un ensemble de n tubes cylindriques élémentaires de section fixe. L'aire de chaque tube élémentaire étant égale à celle de la section de conduit vocal correspondante (voir figure III.4). Il existe ensuite deux approches pour la simulation:

- l'approche acoustique à travers un modèle à réflexion: on parlera ici de simulation temporelle.
- l'approche par analogie acoustico-électrique où chaque tube est remplacé par un circuit électrique équivalent: on parlera ici de simulation fréquentielle.

Sans rentrer dans les détails, nous donnons ci-dessous les grandes lignes de la simulation dans le cas de sons voisés et non voisés.

III.1.2 Modèle à réflexion

Dans un premier temps, nous considérons un tube uniforme de section constante, fermé à une extrémité (au niveau des cordes vocales) et ouvert à l'autre extrémité (au niveau des lèvres).

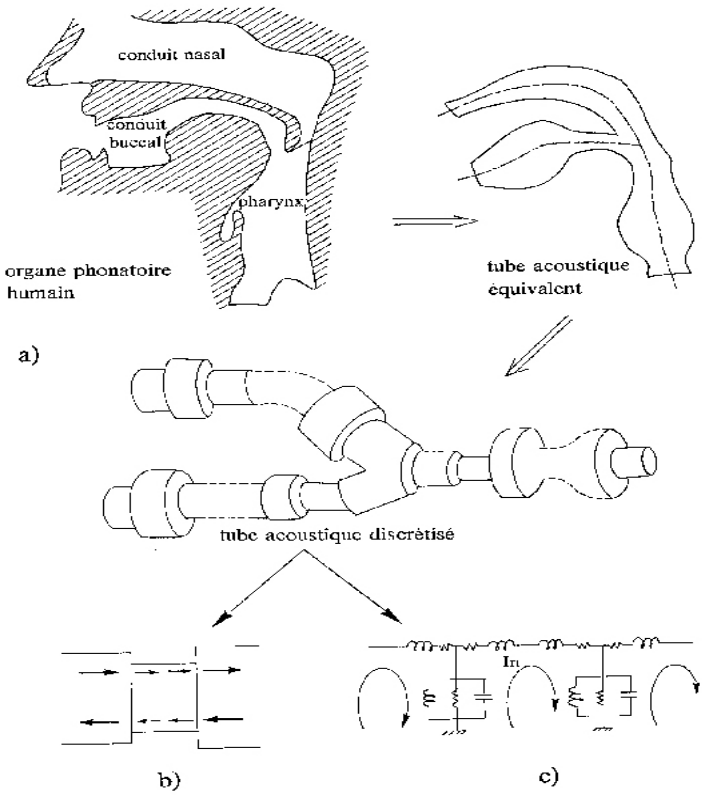


FIG. III.4 – *Modélisation du conduit vocal (d'après [18]): a) Le conduit vocal, b) Le modéle acoustique, c) analogie par cicuit électrique (analogie acoustico-électrique)*

Cette situation est bien évidemment simpliste mais permet de donner une solution initiale à ce problème.

Dans un tel tube, l'aire de la section du tube $A(x)$ est constante et l'on peut donc utiliser les équations de Webster données ci-dessus. La solution de ce problème est bien connu et on sait que les solutions sont du type:

$$p(x,t) = p_+(x-ct) + p_-(x+ct) \quad \text{(III.32)}$$
$$v(x,t) = v_+(x-ct) + v_-(x+ct) \quad \text{(III.33)}$$

où v_+ et v^- représentent les ondes incidentes et réfléchies.

Les solutions sont obtenus en se fixant des conditions aux limites (i.e. à la glotte et au niveau des lèvres). En prenant les conditions aux limites suivantes:

- Au niveau de la glotte ($x = 0$): $u(0,t) = U_g(\Omega)\exp(j\Omega t)$ qui représente une excitation périodique d'amplitude U_g à la pulsation Ω
- Au niveau des lèvres: $p(l,t) = 0$ ce qui correspond à un noeud de pression au niveau des lèvres (ce qui est valable pour un tube ouvert)

On peut montrer (v. [18], p55) que la fonction de transfert (rapport du débit volumique à la glotte au débit volumique au lèvres) s'écrit

$$V = \frac{1}{\cos(\Omega l/c)} \quad \text{(III.34)}$$

Pour des valeurs classiques moyennes du conduit vocal (l = 17,5 cm, et c = 350 m/s), les pôles de cette fonction de transfert, que l'on appelle formants auront pour valeurs 500 Hz, 1500 Hz, 2500 Hz etc...

Pour pouvoir modéliser des formes plus variées du conduit vocal, il sera nécessaire d'abouter plusieurs tubes élémentaires de section constante et d'écrire les équations de propagation à la jonction de deux tubes de sections différentes.

III.1.3 Analogie Acoustico-électrique

La résolution peut être effectuée soit comme suggéré ci-dessus en utilisant les équations de propagation à la jonction de deux tubes de section constantes soit utiliser l'analogie acoustico-électrique et résoudre ainsi le problème par modélisation par ligne électrique.

L'analogie acoustico-électrique est en fait possible en raison des équations de Webster qui relient le débit volumique d'air à la pression. On peut en effet faire le parallèle entre ces équations et les relations qui existent entre la tension $v(t)$ et le courant $i(t)$ en présence d'un inductance L et d'une capacité C:

$$-\frac{\partial v}{\partial x} = L\frac{\partial i}{\partial t} \quad \text{(III.35)}$$
$$-\frac{\partial i}{\partial x} = C\frac{\partial v}{\partial t} \quad \text{(III.36)}$$

les analogies suivantes peuvent ainsi être faites:

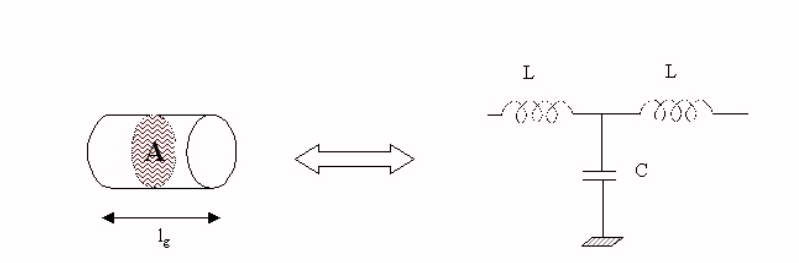


FIG. III.5 – *Analogie Acoustico-électrique: équivalence entre un tube cylindrique de section constante avec paroi rigide et un circuit électrique simple ($L = \frac{1}{A}$ et $C = \frac{A}{\rho c^2}$)*

Quantités acoustiques	**Quantités électriques analogues**
p: pression	i: tension
u: débit volumique	i: courant
ρ/A: inductance acoustique	L: inductance
$A/\rho c^2$ capacité acoustique	C: capacité

Ainsi, un tube acoustique sans perte est équivalent à un circuit électrique simple (voir figure III.5):

En utilisant l'analogie acoustico-électrique, on pourra représenter la pression et le débit volumique à la glotte (respectivement P_g et U_G) en fonction de la pression et du débit volumique aux lèvres (respectivement P_e et U_e) sous la forme:

$$\begin{pmatrix} P_G \\ U_G \end{pmatrix} = T_0 \times T_1 \times \ldots \times T_n \begin{pmatrix} P_e \\ U_e \end{pmatrix} \tag{III.37}$$

où T_n est la fonction de transfert correspondant au quadripole équivalent du n^{ieme} tube acoustique élementaire en partant des lèvres. On parle ainsi communément de simulation fréquentielle pour cette approche. Nous ne détaillerons pas dans ce document cette approche et le lecteur intéressé pourra consulter [18, 88, 89]

III.2 Modélisation des sources vocales

III.2.1 Le larynx

La modélisation de la source vocale (au niveau du larynx) vise soit à reproduire l'onde de débit glottique à travers des modèles géométriques soit à reproduire l'onde de débit glottique à travers un modèle physique.

Dans le cadre de la première approche, un nombre important de modèles ont été proposés. Un grand nombre d'entre eux génère l'onde de débit glottique dans le domaine temporel à l'aide d'un nombre limité de paramètres. On pourra voir figure III.6 les formes d'ondes obtenues par différents modèles dont une étude comparative est donnée dans ([36]).

Pour donner un exemple (le *LF model*) possède 5 principaux paramètres: La fréquence fondamentale, $F0$ et 4 paramètres de forme:

- Ee
- Ei
- te

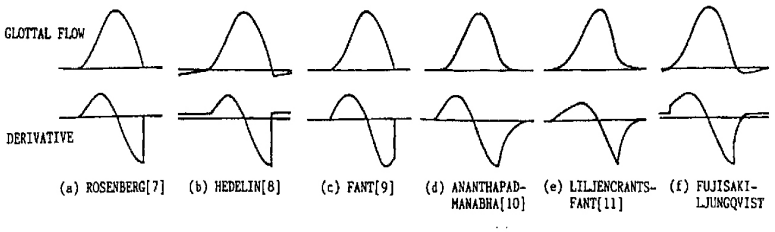


FIG. III.6 – *Formes d'ondes de débit glottique et de leur dérivée pour différents modèles [36])*

– tc

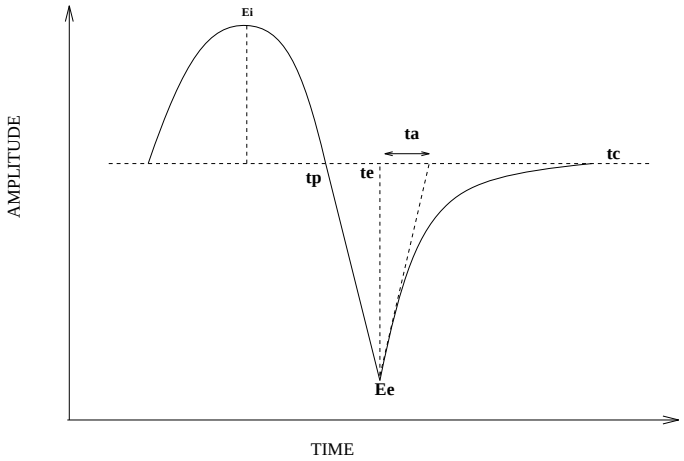


FIG. III.7 – *Modèle LF de source glottale*

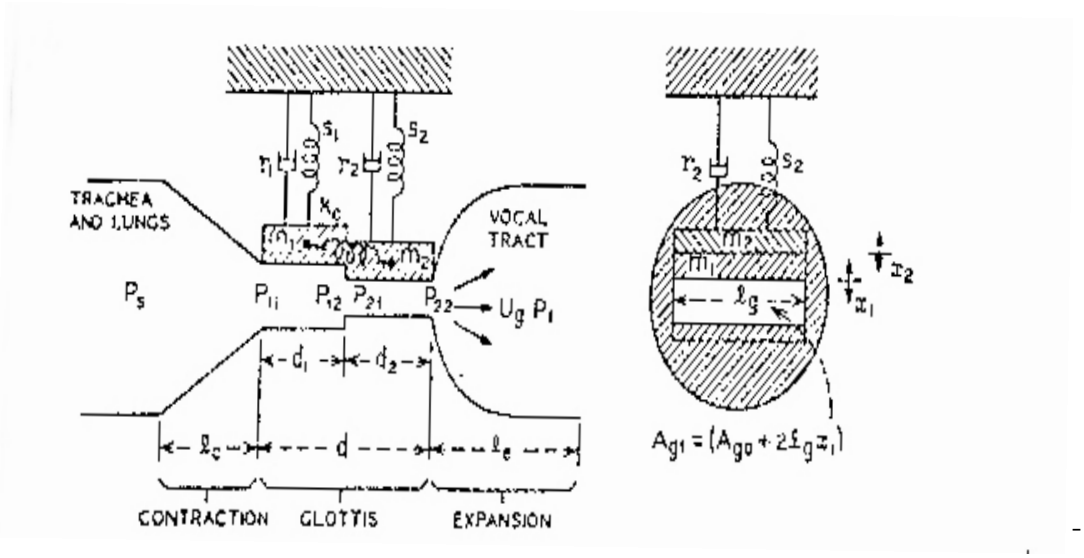
Un certain nombre de modèles articulatoires ont été développés. L'idée la plus communément répandue est de représenter les cordes vocales comme un ensemble de masses reliées entre elles par des ressorts. Le nombre de masses varient considérablement entre les modèles (1 masse pour le premier modèle de Flanagan et Landgrad (1968) jusqu'à N masses pour le modèle de Titze ([18],pp 38).

Le modèle à deux masses introduit par Flanagan et Ishizaka en 1975 ([34]) est probablement le modèle qui a connu (et connaît encore) le plus grand succès en raison de sa simplicité mais aussi bien évidemment en raison de sa bonne représentation des principaux phénomènes de vibration des cordes vocales.

Dans le modèle original, les cordes vocales sont supposées symétriques. Un schéma du modèle à deux masses est donné dans la figure III.8

Les cordes vocales sont donc séparées en profondeur (ou épaisseur) en une partie supérieure et une partie inférieure. Chaque partie est constitué d'un système oscillant possédant une masse (notée m), une raideur (modélisée par un ressort s), et un amortissement (noté r). Les deux masses ne peuvent se déplacer que latéralement (correspondant à des déplacements x_1 et x_2) et sont reliées entre elles par un ressort de raideur k_c. Les autres éléments du modèle sont:

– l_g: la longueur effective des cordes vocales (ou encore de la glotte).
– d_1 et d_2 : l'épaisseur respective des masses m_1 et m_2
– s_1 et s_2: les ressorts respectifs rattachés aux masses m_1 et m_2



-1.5

FIG. III.8 – *Modèle à 2 masses de Flanagan ([34])*

- r_1 et r_2: les résistances respectives aux deux masses m_1 et m_2 représentant les pertes visqueuses.
- A_{g01} et A_{g02}: les aires respectives entre les masses m_1 et m_2 de la glotte lorsque les deux masses sont au repos
- U_g: la vitesse volumique moyenne d'air à travers la glotte

A partir des aires au repos entre les masses m_1 et m_2, il est assez simple d'en déduire l'aire située entre les deux masses en fonction du déplacement x_1 et x_2 de ces masses:

$$A_{g1} = A_{g01} + 2l_g x_1 \quad \text{(III.38)}$$
$$A_{g2} = A_{g02} + 2l_g x_2 \quad \text{(III.39)}$$

L'obtention du flux d'air à la sortie de la glotte s'obtient en résolvant des équations linéaires reliant la différence de pression au débit d'air. De façon générale, le débit U_g satisfait l'équation suivante (si on ne considère pas de sources de bruit à l'intérieur des cordes vocales):

$$R_{tot}U_g + L_{tot}\frac{du_g}{dt} = p_s - p_1 \quad \text{(III.40)}$$

où p_s est la pression subglottique (la pression provenant des poumons), p_1 est la pression à la sortie de la glotte, R_{tot} et L_{tot} étant respectivement la résistance et l'inductance (quasi-stationnaires) représentant la contraction, la glotte et l'expansion. En pratique, il sera nécessaire d'écrire et de résoudre ces équations pour chaque élément (contraction, première partie de la glotte, seconde partie de la glotte et expansion, voir figure III.9, car pour chaque partie des phénomènes physiques différents entrent en jeu et donnent lieu à des hypothèses différentes pour leur résolution). Ces équations seront ensuite résolues numériquement en discrétisant les éléments différentiels.

Nous ne donnerons pas plus de détails dans ce cours de ce modèle, en précisant toutefois que ce modèle est particulièrement réaliste tout en restant simple et qu'il permet également de mettre en évidence les phénomènes de couplage entre la source vocale et le conduit vocal.

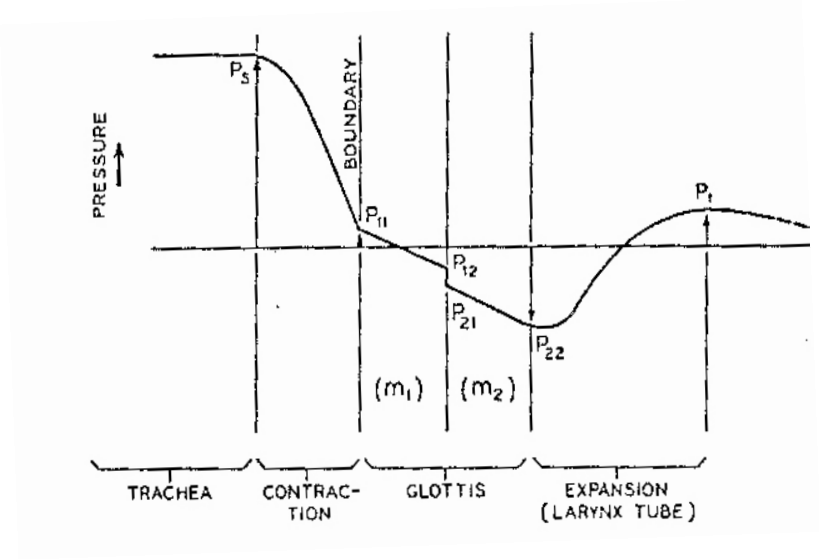


FIG. III.9 – *Les différentes régions du modèle à deux masses ([34])*

III.2.2 Les sources de bruit (ou frication)

Pour les tout premiers synthétiseurs mécaniques, l'air était un élément physique à part entière (comme dans la machine de Von kempelen où l'on trouve un soufflet). Lorsque les premiers synthétiseurs électriques sont apparus, ils furent développés en utilisant la théorie de l'acoustique linéaire telle qu'elle a été décrite ci-dessus. Hors, dans ce cadre, il est nécessaire de définir un modèle particulier pour les sons non voisés tels que les fricatives puisque ces sons sont le résultat d'un phénomène non-linéaire. Il existe assez peu de modèles et ils sont pour la plupart basés sur une idée assez simple que la source de bruit peut être représentée comme une fluctuation aléatoire des grandeurs aérodynamiques (pression ou vitesse volumique).

L'un des modèles qui a été assez utilisé est celui de Flanagan et Cherry ([32],p253-259)) en raison de sa simplicité mais aussi en raison des résultats plutôt satisfaisant que l'on obtient malgré la complexité des mécanismes entrant en jeu lors de la production d'un bruit de friction dans le conduit vocal.

Ce modèle est basé sur des observations expérimentales ([67]) qui ont permis de relier la pression de la source de bruit au carré du nombre de Reynolds définit par:

$$R_e = \frac{\rho_0 du}{\mu} = \frac{\rho_0 dU}{\mu A} \tag{III.41}$$

où

- ρ_0 est la densité de l'air
- u est la vitesse des particules d'air (qui est relié à la vitesse volumique de l'air par $u = U/A$)
- μ est le coefficient de viscosité
- d est la largeur de la constriction

Ainsi, la pression de la source est proportionnelle au carré de la vitesse d'air à partir d'un seuil déterminé expérimentalement. Cette source possède une impédence interne R_n qui est elle aussi déterminée expérimentalement et suit egalement une loi en fonction du carré de la vistesse volumique. Pour la synthèse, il suffira de mettre cette source en série entre deux circuits électriques équivalents :

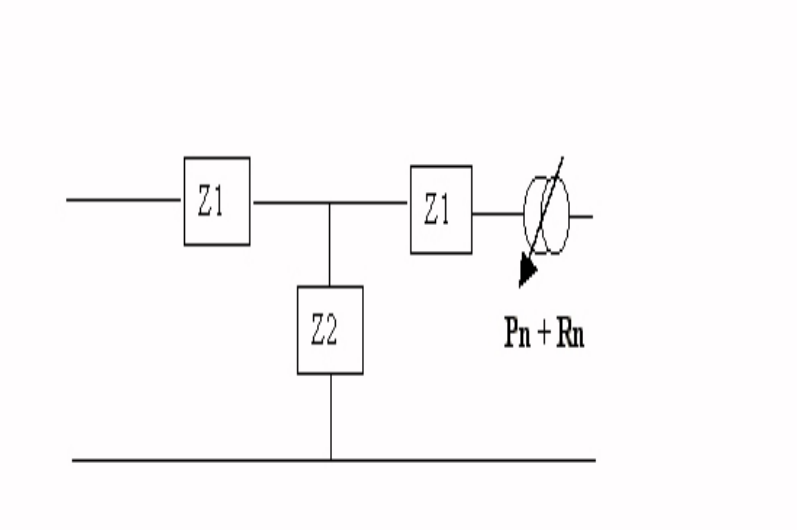


FIG. III.10 – *Circuit életrique équivalent modifié avec l'adjonction d'une source de bruit (pression aléatoire), P_n et son impédence interne R_n*

Ce modèle simple ne modélise pas vraiment les mécanismes physiques de production de sources non-voisées dans le conduit vocal mais la théorie aéroacoustique de la génération d'un son dans un tuyau permet d'expliquer ce modèle. En effet, cette théorie montre que la pression acoustique varie comme le carré de la vitesse d'air. Cependant, il existe tout de même un certain nombre d'inconvénients dans ce modèle. Il suppose en effet que:

- les fluctuations de pression sont supposées localisées en un point précis du conduit vocal
- l'effet de la géométrie du conduit vocal sur les caractéristiques de la source est décrit par un seul spectre de source
- la source est supposée être localisée soit au niveau de la principale constriction soit en un point précis arbitraire en avant de la constriction.

Cependant, le principal défaut d'un tel modèle est qu'il ne tient pas compte des caractéristiques physiques de production. Un certain nombre de travaux ([65, 85] se dirigent ainsi vers l'utilisation de la théorie de l'aéroacoustique pour en déduire des modèles plus réalistes de source non-voisée dans le conduit vocal. La théorie de l'aéroacoustique repose sur les travaux précurseurs de Lighthill ([58])

Sinder [85] a récemment proposé un nouveau modèle de source pour les fricatives en intégrant une modélisation des mouvements des tourbillons d'air se créant au niveau d'une constriction et qui sont responsables de la création d'une source sonore. Ce modèle s'appuie sur trois modules principaux (voir figure III.11):

- Un modèle de Jet d'air appelé, "*Jet model*", qui décrit la formation, la convection des tourbillons en incluant leur puissance, leur vitesse et leur trajectoire.
- Un modèle de flux d'air moyen décrivant la direction de ce flux irrotationnel.
- Un modèle de propagation acoustique, qui résout les équations de propagation des ondes acoustiques sachant qu'une description des sources acoustiques est donnée.

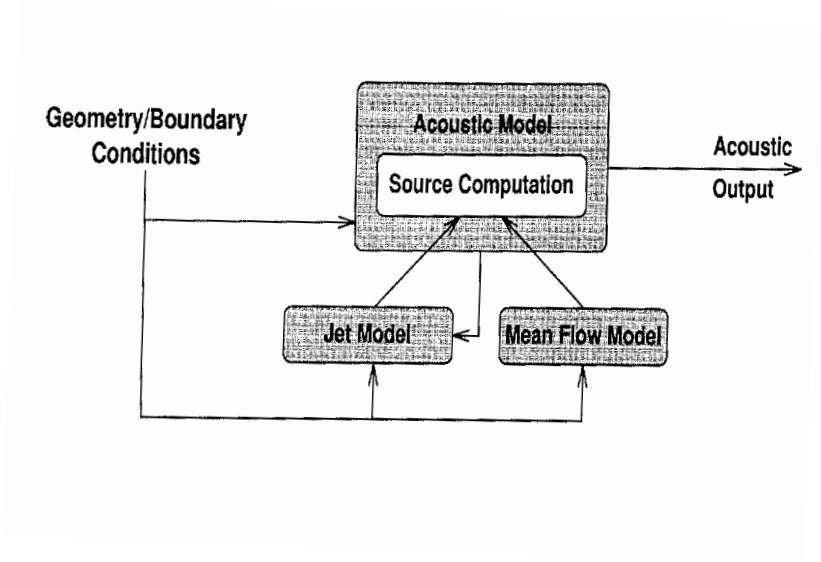


FIG. III.11 – *Schéma illustrant le modèle utilisant la théorie de l'aéroacoustique (d'après [85])*

Ce modèle, plus proche de la physique, permet de générer des fricatives sans connaissance a priori sur la force, la localisation et le spectre des sources de bruit. Le lecteur intéressé pourra consulter ([85]).

III.2.3 Le conduit vocal

Il existe de même un grand nombre de modèles pour le conduit vocal et pour les articulateurs (langues, machoires, etc...). Les modèles articulatoires peuvent être statiques ou dynamiques, descriptifs ou fonctionnels ([83]). Les modèles dynamiques (par exemple celui de Henke) est contrôlé par des gestes ou des cibles articulatoires. Un tel modèle est donc régit par des équations de mouvement des articulateurs. Une autre famille de modèles est basée sur une modélisation par composantes linéaires: les différentes formes du conduit vocal (obtenues par imagerie médicale, Rayons X ou IRM) sont décrites comme la somme de composantes linéaires. Le modèle de S. Maeda en est un exemple et est décrit plus en détail ci-dessous ([62]).

Le Modèle de S. Maeda

Le modèle de Maeda est un modèle intéressant dans le sens où il est obtenu à l'aide de méthodes statistiques sur des données d'observation. Ces observations sont représentées à l'aide d'une grille semi-polaire (voir sur la partie gauche de la figure III.3). Le principe de base de ce modèle est d'appliquer une analyse en composantes linéaires sur les vecteurs d'observation (en pratique on utilisera en fait des observations centrées et normées c'est à dire de variance unitaire). Ainsi, les formes observées sont décrites sous la forme:

$$X = \sum_{i}^{p} a_i y_i \tag{III.42}$$

où X est le vecteur des observations (distances pour chaque point entre le contour inférieur et le contour supérieur du modèle), a_i,$i = 1..p$ sont les coefficients caractérisant chaque paramètre y_i du modèle. Le principe de base consiste à calculer la composante lié à un paramètre puis à la

soustraire avant de calculer les composantes liés à une autre région du conduit vocal. L'analyse débute par le paramètre de mâchoire qui est plus facilement observable sur les données. Cette approche a abouti à un modèle à sept paramètres.

Le modèle de Flanagan-Ishiza

Finalement il existe une autre famille de modèles, plus simples, qui sont les modèles descriptifs statiques comme les modèles de Coker, Mermelstein et Flanagan et Ishizaka.

Ces modèles sont également régis par un nombre limité de paramètres mais qui s'avèrent très utiles pour l'inversion acoustico-articulatoire. L'un de ces modèles est celui de Flanagan Ishizaka donné figure III.12. Ce modèle simple représente la fonction d'aire du conduit vocal sous la forme de l'équation suivante:

$$A(x) = (\frac{A_b + A_c}{2}) - (\frac{A_b - A_c}{2})cos[\pi(\frac{X_c - X}{l_b})], \qquad X \leq X_c \tag{III.43}$$

$$= (\frac{A_f + A_c}{2}) - (\frac{A_f - A_c}{2})cos\pi[0.4 + 0.6(\frac{X - X_c}{l_f})](\frac{X - X_c}{l_f}), \qquad X > X_c \tag{III.44}$$

avec:

$$A_c \leq A_b; A_f \tag{III.45}$$

$$A_c > 0 \tag{III.46}$$

$$13 \leq L \leq 21cm \tag{III.47}$$

$$L/10 < X_c < 9L/10 \tag{III.48}$$

et où

- A_b est l'aire de la cavité arrière
- A_f est l'aire de la cavité avant
- A_c est l'aire au niveau de la principale constriction du conduit vocal
- A_f est l'aire au niveau des lèvres
- X_c est la position de la principale constriction
- L est la longueur du conduit vocal
- $l_b = \frac{8L}{17}$ est la distance entre la principale constriction et la position de la section d'aire maximale de la cavité arrière.
- $l_b = \frac{7L}{17}$ est la distance entre la principale constriction et la position de la section d'aire maximale de la cavité avant.

Modèles déductifs[1]

Les modèles articulatoires permettent de simuler l'appareil de production de parole et de tester l'importance de tel ou tel paramètre. On peut ainsi étudier le rôle de la cavité du larynx : par exemple, est ce que l'absence de cavité du pharynx chez le singe, ce qui peut limiter les capacités acoustiques du conduit vocal, est à l'origine du fait qu'il ne parle pas ([72]) ? Avec cette hypothèse, la descente du pharynx, dont l'origine est très discutée, aurait été fondamentale pour le développement du système de communication parlée ? On aimerait aussi pouvoir étudier

1. Ce paragraphe a été rédigé par René Carré

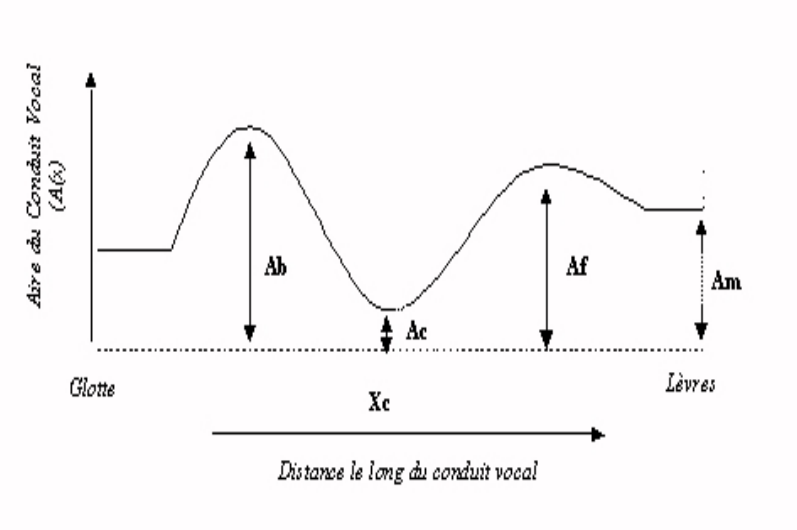


FIG. III.12 – *Modèle de Flanagan-Ishizaka (d'après [35])*

la question suivante : Est-ce que le conduit vocal humain est aujourd'hui optimal pour répondre aux besoins de communication ? S'il est optimal, est-ce par hasard ou bien est-ce le résultat d'une évolution pilotée par les besoins de communication ? Dans ce cas, on devrait pouvoir prédire les phases d'évolution à partir d'une situation non optimale donnée (de type ancestral). Pour cette dernière étude, l'approche déductive qui permet d'aller au-delà des observations et des données accumulées sur les langues et sur les formes du conduit vocal peut être plus productive. L'étude de la communication parlée, comme de tout phénomène physique, peut donc être abordée en suivant deux approches complémentaires ([5]):

- La première consiste à accumuler, et à interpréter des données : sur le signal de parole, la fonction d'aire du conduit vocal, l'appareil vocal, les systèmes phonologiques, Ces données peuvent être représentées par des modèles qui simplifient la représentation des données mais de tels modèles n 'explique pas les phénomènes sous-jacents : une donnée est expliquée par les autres données.
- La deuxième approche est déductive : les données sont, dans la mesure du possible, observées de l'extérieur. Par exemple, le signal de parole n'est plus analysé en termes de fréquences de formants mais selon l'appareil de production (l'appareil vocal humain) qui a généré ce signal ou/et l'appareil de perception qui analyse ce signal. Une telle approche a conduit Liljencrants et Lindblom ([59]) à déduire avec un certain succès, à partir de caractéristiques de l'appareil de production associées à un critère de contraste perceptif, des systèmes vocaliques expliquant les inventaires des langues du monde ([51]). En poursuivant ce type de démarche jusqu' à l'extrême, on peut entreprendre une étude sur l'origine des caractéristiques de l'appareil de production (et de perception) ? Ces caractéristiques ne sont-elles pas le résultat d'une évolution visant à disposer du meilleur système de communication acoustique possible (en partant du principe que la communication est essentielle pour préserver l' existence de l'être humain) ?

On peut en effet montrer qu'en exploitant les caractéristiques acoustiques d'un tube de 18 cm de long (longueur du conduit vocal humain) selon les seuls critères d'efficacité et de contraste

acoustique, le modèle obtenu est comparable avec le système de production de parole chez l'homme ([79]).

III.3 Inversion acoustico-articulatoire

En associant un modèle de glotte à un modèle de conduit vocal, il est possible de synthétiser des sons. Les paramètres pouvant être réglés manuellement, des exemples de synthèse de très bonne qualité peuvent être obtenus.

Cependant, il existe de nombreux intérêts à pouvoir estimer automatiquement ces paramètres à partir du signal de parole. On parle ici d'*inversion acoustico-articulatoire* puisqu'à partir d'un signal acoustique on recherche les différents paramètres articulatoires qui permettraient de générer un son approchant au mieux le signal original. Le potentiel d'une telle inversion est très grande puisqu'elle permettrait d'envisager des applications dans de nombreux domaines du traitement de la parole tel que le codage bas débit, la reconnaissance à partir des trajectoires articulatoires ou utilisant la représentation articulatoire comme une information supplémentaire et bien sur la synthèse où les paramètres qui sont liés à la physiologie permettent un contrôle aisé et intuitif de la qualité d'une voix et de sa transformation.

III.3.1 Approches à l'aide de tables

Ainsi de nombreuses approches ont été proposées pour réaliser cette inversion. Ce problème est cependant loin d'être trivial puisque qu'il n'existe pas une solution unique, c'est à dire qu'il existe plusieurs formes du conduit vocal qui peuvent donner lieu à une fonction de transfert donnée. L'une des approches qui a été particulièrement étudiée consiste à construire des tables (ou *codebook*) qui associent un ensemble de paramètres articulatoires à une représentation spectrale du signal acoustique correspondant. Lors de l'estimation, le signal de parole original est comparé (par exemple à l'aide d'une distance spectrale pondérée perceptuellement) à chaque élément de la table. L'ensemble des paramètres articulatoires produisant la parole synthétique la plus proche du signal original est alors sélectionné. Une telle approche peut être élargie en multipliant les tables suivant la nature des signaux (on pourra par exemple définir une table pour les sons voisés et une table pour les sons non-voisés). En pratique, une inversion par table nécessite des étapes supplémentaires pour soit affiner l'estimation (les tables ne pouvant représenter tout l'espace articulatoire) soit pour lisser les trajectoires et lever l'ambiguité liée à la non-unicité des solutions. Cette seconde étape peut-être réalisée en minimisant la distorsion entre le signal original et le signal de synthèse (La recherche du minimum peut être par exemple réalisée par un algorithme du gradient sur les paramètres du modèles articulatoire). Un exemple d'un système d'inversion articulatoire à l'aide du modèle de Flanagan-Ishizaka est donné figure III.13, ou le lissage des trajectoires est réalisé à l'aide d'un algorithme de DTW (Dynamic Time Warping).

Les tables (ou codebook) utilisées pour l'estimation initiale de la forme du conduit vocal doivent couvrir l'espace des paramètres articulatoires d'un locuteur. De plus, l'échantillonnage de cet espace doit être particulièrement fin pour pouvoir garantir que l'estimation initiale est proche des valeurs optimales. Ainsi, de telles tables requièrent un très grand nombre de paires (paramètres du modèle de production, paramètres spectraux du signal de parole synthétisé correspondant) et la recherche exhaustive dans ces tables devient rapidement fastidieuse. En pratique, il est ainsi souhaitable d'optimiser cette recherche. Une approche consiste à regrouper les entrées de la table en un certain nombre de cluster. Le regroupement peut s'effectuer à partir des caractéristiques acoustiques des signaux synthétiques. Ainsi, chaque élément d'un cluster est

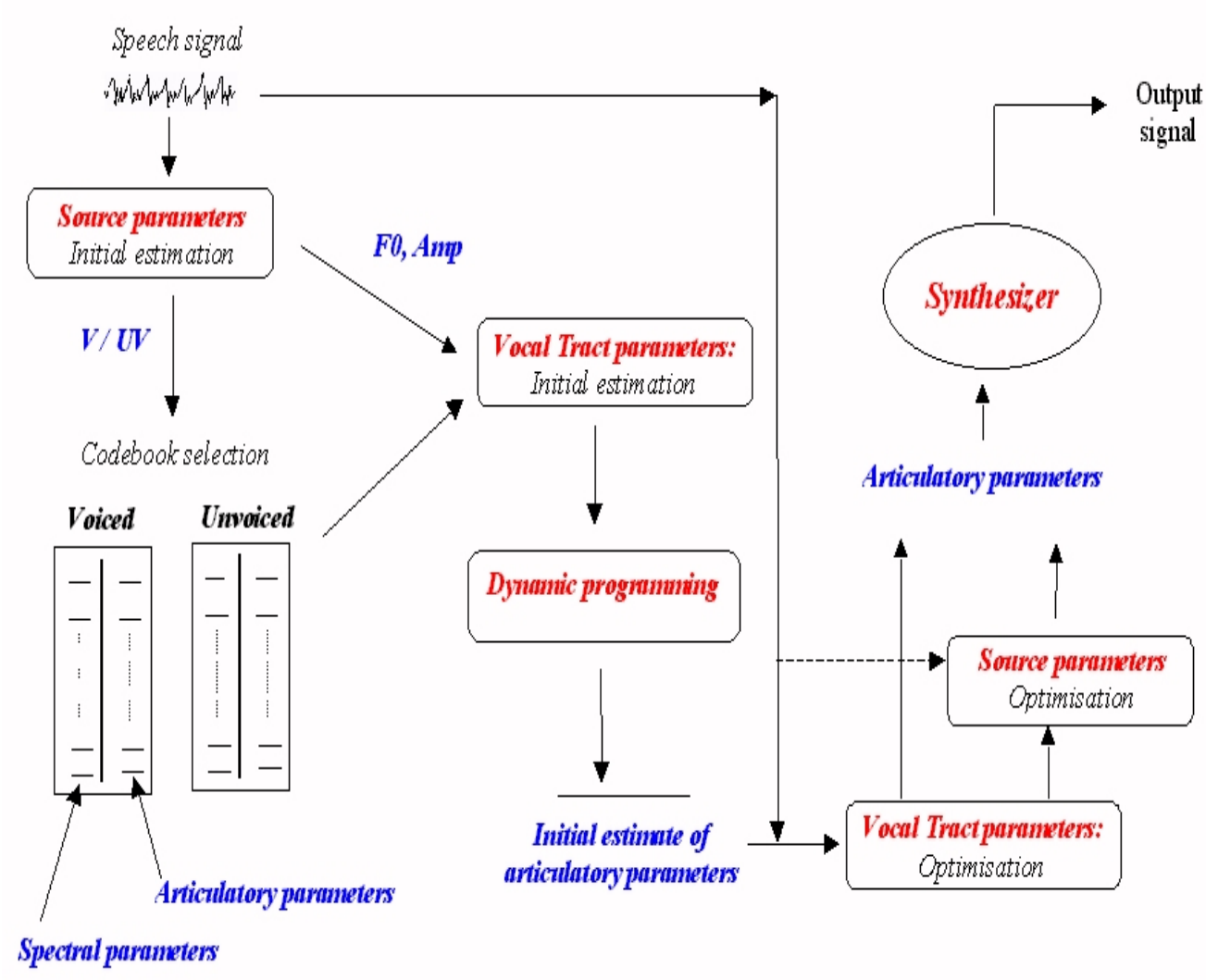


FIG. III.13 – *Principes d'analyse/synthèse à l'aide d'un modèle articulatoire (d'après [40])*

relié à l'ensemble des formes de conduit vocal qui produisent des signaux aux caractéristiques proches. La recherche est ainsi grandement accélérée et ce quelque soit la taille de la table ([82]).

III.3.2 Autres approches

Il existe un très grand nombre de méthodes pour associer les paramètres articulatoires aux paramètres acoustiques. On pense notamment aux approches par réseaux de neurones, aux approches statistiques (Modèles de markov cachés, réseaux bayésiens, etc....), aux approches utilisant des fonctions élémentaires (*basis functions*) qui peuvent d'ailleurs être vues comme une classe particulières des réseaux neuronaux ([73, 83]) ou encore aux méthodes utilisant des algorithmes génétiques pour retrouver les mouvements articulatoires à partir des fréquences des 2 premiers formants ([65]).

Une autre méthode largement utilisée est l'approche par *regression non linéaire* qui est succinctement décrite ci-dessous. Cette méthode consiste à trouver une fonction g, telle que $y = g(x)$ où y est le vecteur des paramètres articulatoires, x le vecteur des paramètres acoustiques et g les coefficients de la forme polynomiale. Ces coefficients sont estimés en minimisant l'erreur quadratique moyenne sur un corpus d'apprentissage de mesures de x et y.

Soit x_i,$i = 1$,...n_x, y_i,$i = 1..n_y$ et g_i,$i = 1..n_y$ les composantes respectives des vecteurs x, y et g. En écrivant un développement en série de Taylor au second ordre, on peut écrire:

$$\hat{y_i} = y_i^0 + \sum_{k=1}^{n_x} \delta_{ik} x_k + \sum_{k=1}^{n_x} \sum_{j=1}^{n_x} \gamma_{ijk} x_j x_k \tag{III.49}$$

où y_i^0 est la valeur (inconnue) de g_i en x^0.

De façon plus condensée, on peut écrire:

$$\hat{y_i} = y_i^0 + \sum_{k=1}^{n_x} b_{ik} \tilde{x_k} \tag{III.50}$$

où b_{ik} représentent les coefficients δ_{ik} et γ_{ijk} et $\tilde{x_k}$ regroupe les termes linéaires et quadratiques des composantes de $\mathbf{x}$.

On peut réécrire sous forme matricielle, l'équation précédente:

$$\hat{\mathbf{y}} = \mathbf{y^0} + B\tilde{\mathbf{x}} \tag{III.51}$$

La solution de ce problème est obtenu en minimisant l'erreur quadratique E entre $\mathbf{y_i}$ et $\hat{\mathbf{y}}_\mathbf{i}$ ce qui permettra d'obtenir $\mathbf{y^0}$ et $\mathbf{B}$:

$$E = \sum_{i=1}^{M} \|\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}| \tag{III.52}$$

III.4 Modélisation 3D pour la synthèse audio-visuelle

Il est un autre domaine où la représentation possède un formidable potentiel: celui de la synthèse audio-visuelle (encore appelée animation de visages parlants). En effet, une bonne modélisation tri-dimensionnelle des articulateurs (notamment ceux visibles de l'extérieur) permettent de développer des visages parlants. L'apport de la modalité image s'avère importante pour:

– l'intelligibité de la synthèse de parole (voir ref dans Guiard-Marigny96).

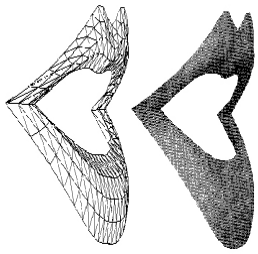


FIG. III.14 – *Modèle 3D de lèvres (d'après [93])*

– le naturel et la convivialité de l'interface ainsi crée

Pour la parole, il est clair que ce sont les lèvres, la mâchoire et la langue qui reçoivent le plus grand intérêt. Pour construire ces modèles tout en limitant le nombre de paramètres nécessaires pour leur contrôle, il est souvent fait appel aux techniques d'analyse en composantes linéaires (ACL) ou d'analyse en composantes principales (ACP), comme cela a été fait pour le modèle de S. Maeda.

Pour ces modèles, il est nécessaire d'avoir des données en quantité suffisante. Pour les lèvres, on pourra pour obtenir des données facilement extractibles, utiliser par exemple un rouge à lèvres de couleur bleue. Les premiers modèles furent développés en 2D et ont consisté à définir un ensemble de base de contours de lèvres. Dans [12], un ensemble de 22 formes de bases, appellées "visèmes", ont été utilisés. A l'aide de ces visèmes, il est possible de représenter les contours à l'aide d'équations polynomiales simples en fonction d'un nombre limité de paramètres. Par exemple, Guiard propose 3 paramètres pour le contrôle des contours de lèvres:

– La hauteur (largeur d'ouverture des lèvres)
– La largeur des lèvres
– le degré de protrusion.

Pour un modèle 3D 2 paramètres supplémentaires ont été ajoutés , soit:

– La protrusion de la lèvre supérieure
– La protrusion de la lèvre inférieure

Afin de rendre le volume des lèvres, trois contours intermédiaires sont identifiés entre le contour interne et externe ce qui a donné lieu à dix équations polynomiales dont les coefficients peuvent être prédit à partir des 5 paramètres donnés ci-dessus. Un exemple de ce modèle est donné sur la figure III.14 avec la structure fine sous jacente.

Après les lèvres, la mâchoire est certainement l'articulateur le plus visible (même si l'on peut prononcer de la parole intelligible sans actionner la mâchoire). Pour obtenir des visages parlants naturels il est ainsi important de pouvoir bien modéliser ses mouvements. La mâchoire possède un nombre de degré de liberté plus faible que les lèvres. En effet, étant un objet rigide non déformable, la mâchoire possède 6 degrés de liberté, que l'on peut visualiser sur la figure ci-dessous. Le modèle est construit à partir d'une grille de 6000 polygones.

La langue est également un articulateur important et peut être utile pour la synthèse audio-visuelle. Néanmoins, son étude est un peu plus complexe puisque la langue n'est pas entièrement visible de l'extérieur. On a ainsi recours à l'imagerie médicale, comme pour l'étude du conduit vocal (rayons X, images IRM).

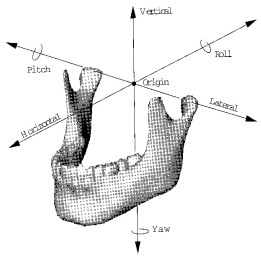


FIG. III.15 – *Modèle 3D de la mâchoire (d'après [93])*

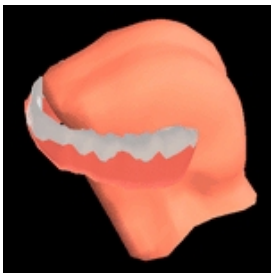


FIG. III.16 – *Modèle 3D de la langue (d'après [71])*

De même que pour les autres articulateurs, un certain nombre de modèles 2D et 3D de la langue ont été proposés. Si certains modèles se fondent sur des modèles à éléments finis, la tendance actuelle est de construire ces modèles à partir d'analyse statistique (ICA,ICP) sur des données réelles.

Le modèle suivant développé par Engwall ([28]) suit cette approche et définit 6 paramètres pour contrôler son modèle de langue:

- JH: L'ouverture de la mâchoire qui mesure la position de la mâchoire par rapport au palais
- TB: Le corps de la langue qui contrôle le mouvement avant-arrière de la langue réalisé en relevant la langue vers le palais tout en contractant dans la région pharyngale (correspond à l'activation du muscle genioglossus).
- TD: Le dos de la langue qui contrôle la courbure du corps de la langue (qui correspond à l'activation du muscle styloglossus)
- TT: représente la position de la pointe de la langue dans un plan vertical (haut-bas)
- TA: représente la position de la pointe de la langue dans un plan horizontal (avance-rétraction).
- TW: largeur de la langue

Un exemple de ce modèle est donné figure III.16:

Le lecteur intéressé pourra consulter le site de P. Badin ([6]) qui contient de plus amples informations et un nombre conséquent de films audio-visuels très démonstratifs. On pourra aussi consulter le chapitre 4 du récent ouvrage sur l'analyse, la synthèse et codage de la parole [64]

Chapitre IV

Reconnaissance de la parole

IV.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 IV.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 IV.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

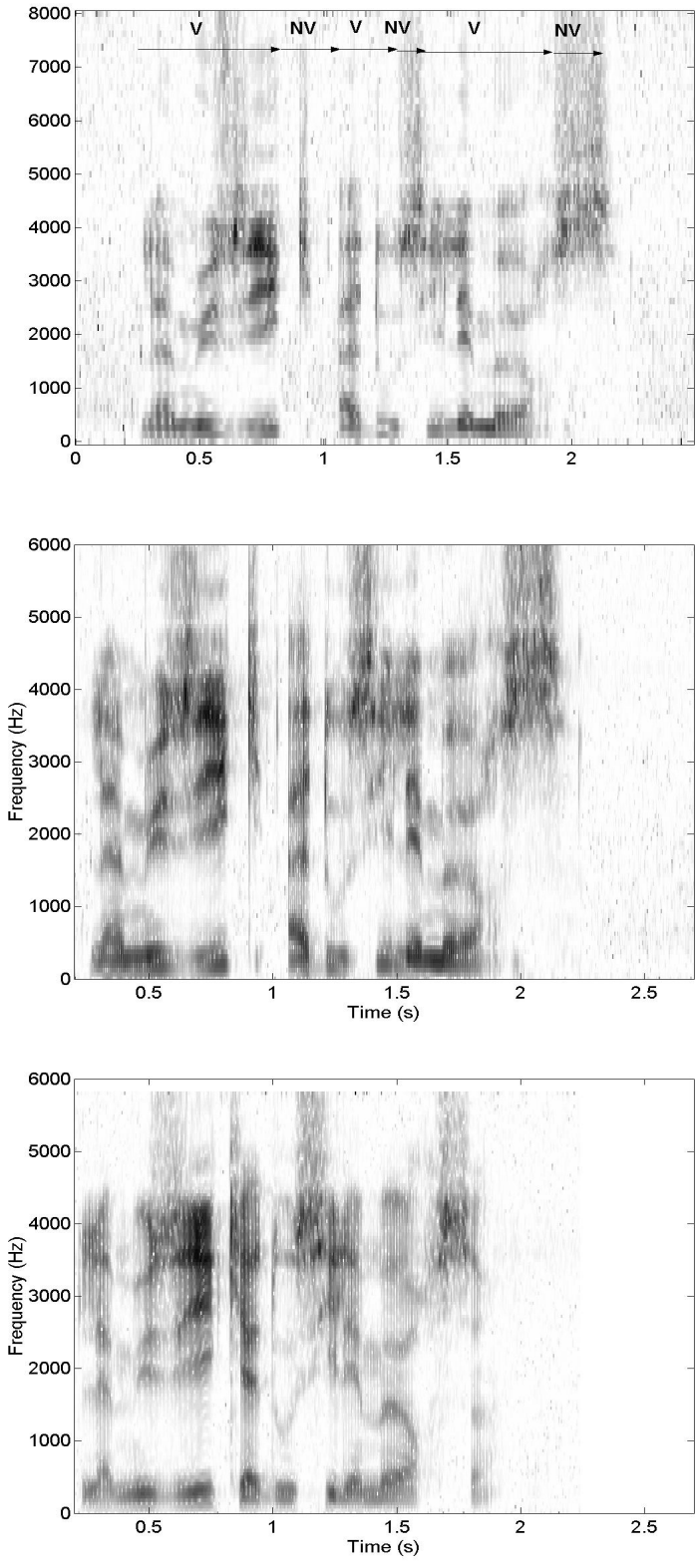


FIG. IV.1 – *La phrase* "La musique adoucit les moeurs" *prononcée trois fois par le même locuteur*

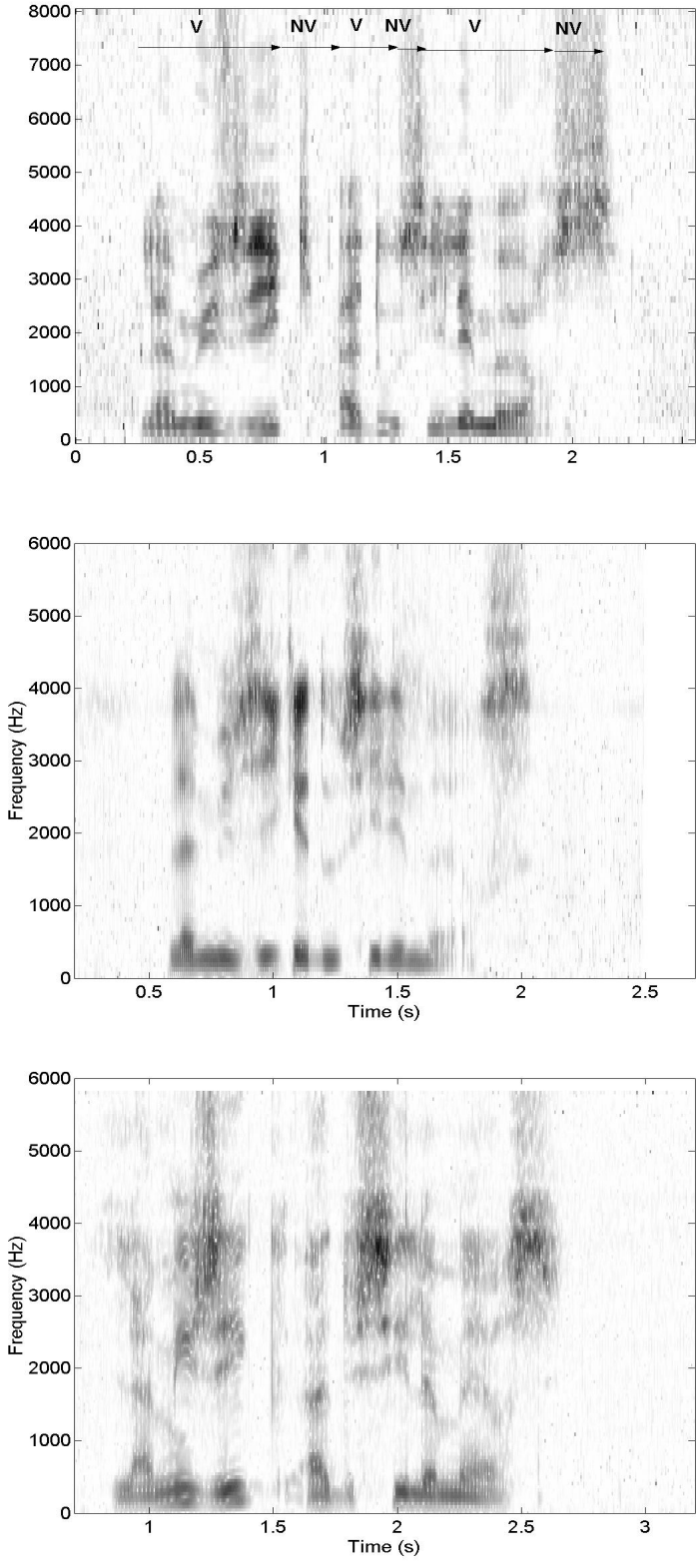


FIG. IV.2 – *La phrase* "La musique adoucit les moeurs" *prononcée par trois locuteurs différents*

et aux conditions d'enregistrement. Les conditions optimales pour la reconnaissance vocale sera un 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 ([15],[38]):

- *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).

IV.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.

IV.2.1 Les approches basées sur les connaissances (ou approches acoustico-phonétiques)

Principes

Cette approche est basée sur les connaissances phonétiques et fait l'hypothèse qu'il existe un nombre finis d'unités phonétiques distinctes dans une langue parlée et que ces unités phonétiques sont grossièrement caractérisées par un ensemble de propriétés visibles (ou analysables) au cours du temps. Même si les propriétés acoustiques de ces unités phonétiques sont fortement variables en fonction des locuteurs mais aussi en fonction des phonèmes adjacents (le fameux phénomène

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

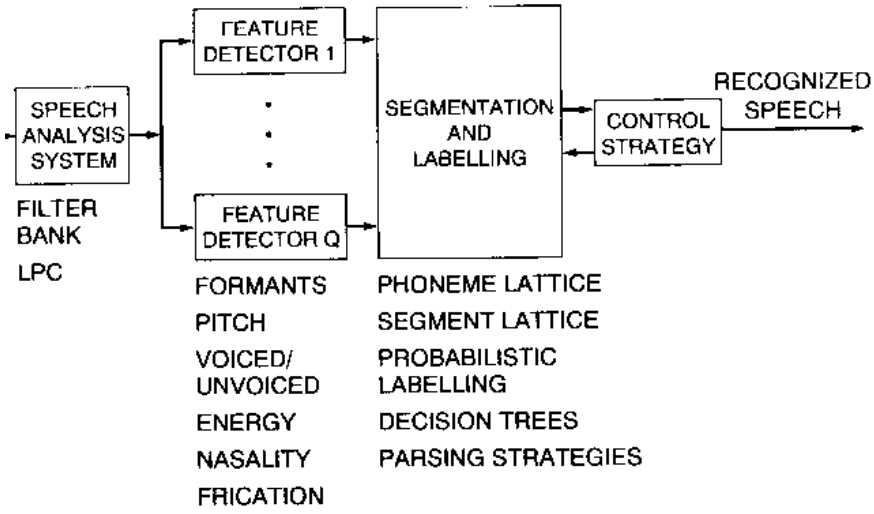


FIG. IV.3 – *Schéma bloc d'un système de reconnaissance vocale par une approche basée sur les connaissances [80]*

de coarticulation), il est supposé que les règles décrivant cette variabilité sont simples et qu'elles peuvent être apprises. La reconnaissance suivant cette approche est ainsi réalisée en 2 étapes:

- Une étape de *Segmentation et Labellisation* qui consiste à segmenter le signal de parole en éléments courts pour lesquels les propriétés acoustiques sont représentatives d'un ou d'une classe de phonème(s), puis à attacher un ou plusieurs label(s) phonétique(s) à chaque segment en fonction de ses propriétés acoustiques
- Une étape *d'identification* qui consiste à déterminer le mot ou (la chaîne de mots) qui a été prononcé en fonction des labels phonétiques produits par la première étape. Cette étape est très importante et n'est pas aussi évidente qu'il n'y parait. En effet, l'étape de segmentation/labellisation fournit en général plusieurs phonèmes candidats pour chaque segment résultant ainsi en un treillis phonétique pour lequel de nombreuses possibilités de mots sont possibles.

La figure IV.3 décrit plus précisément les principales étapes de cette approche. Tout d'abord, le signal de parole $s(n)$ est analysé pour en obtenir une représentation appropriée (par exemple spectrale) de ses caractéristiques. Nous avons ensuite une étape de détection de caractéristiques, l'idée étant ici de convertir la représentation spectrale en un ensemble de caractéristiques regroupant les propriétés acoustiques des différentes unités phonétiques. Par exemple sur l'exemple donné figure IV.3, ces caractéristiques sont les formants, la fréquence fondamentale, le voisement (présence ou absence de voisement), le rapport d'énergie entre les hautes et basses fréquences (paramètre *Energy*), la présence ou l'absence de bruit fricatif (paramètre *Frication*).

Bilan de cette approche

Les intérêts d'une telle approche sont nombreux. Tout d'abord, elle permet de générer des systèmes à vocabulaire illimité puisqu'elle obtient une représentation sous forme de suite de phonèmes. Cette approche est de plus générique et peut être appliquée à d'autres langues (même

si un travail d'analyse important pour chaque nouvelle langue se révèle nécessaire). De plus, cette approche permet d'affiner les connaissances sur la parole aux niveaux production et perception.

Cependant, cette approche connaît en pratique de nombreux problèmes et n'a pas vu la réalisation d'applications effectives. Parmi ces problèmes, on peut citer:

- La nécessité d'avoir des connaissances approfondies des propriétés acoustiques des unités phonétiques et cette connaissance est, bien évidemment, incomplète.
- Le choix des caractéristiques est fait principalement sur des considérations *ad hoc* et est généralement le fruit de l'intuition. Ainsi, les caractéristiques retenues (ou paramètres) ne sont pas nécessairement optimales pour un système de reconnaissance.
- Il n'existe pas de méthode automatique pour régler les différents paramètres du système (i.e. ajuster les seuils de décision, etc.) sur des données labellisées de parole. D'ailleurs, la phonétisation d'un corpus (i.e. l'action de transcrire sous forme phonétique des données enregistrées de parole) n'est pas unique et varie de façon significative suivant les experts linguistiques.

En raison de ces problèmes, l'approche acoustico-phonétique garde des perspectives intéressantes mais nécessiterait des efforts importants de recherche avant de pouvoir l'appliquer avec succès au problème de la reconnaissance de parole.

IV.2.2 Les approches d'intelligence artificielle

Les approches d'intelligence artificielle sont des méthodes hybrides entre les approches acoustico-phonétiques et les approches statistiques dans le sens où elles exploitent des idées des deux concepts. La démarche de l'intelligence artificielle a pour but d'automatiser la procédure de reconnaissance en prenant en compte la façon dont on utilise notre intelligence pour visualiser, analyser et finalement pour prendre une décision sur les caractéristiques acoustiques mesurées. Comme mentionné ci-dessus, les systèmes experts entrent dans cette catégorie. Par exemple, un tel système effectuera une segmentation puis une labellisation en utilisant plus de niveau de connaissance que le seul niveau acoustico-phonétique (en particulier, il utilisera des connaissances lexicales, syntaxique, sémantiques etc..). Les réseaux de neurones ont été souvent rattachés à cette famille pour apprendre les relations entre les unités phonétiques et des paramètres d'entrée. Cependant, l'utilisation des réseaux est plus générale et peut tout aussi bien être rattachée aux approches statistiques.

IV.2.3 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

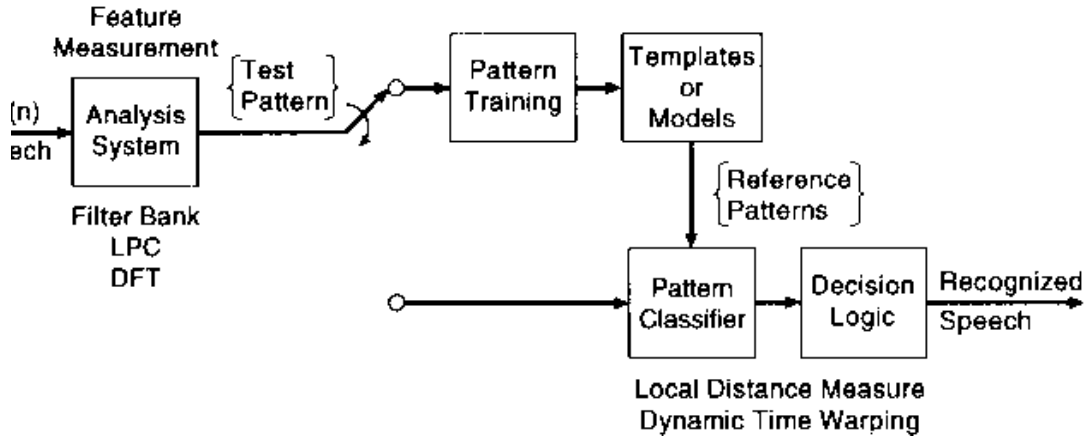


FIG. IV.4 – *Schéma bloc d'un système de reconnaissance vocale par une approche statistique [80]*

élémentaire apprise durant la phase d'apprentissage permet de classifier le signal d'entrée en fonction de ces unités.

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 IV.4 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 IV.8.
- 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.

IV.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 ([38]) :

- 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 IV.5 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 ([43]) 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

FIG. IV.5 – *Tableau récapitulant les valeurs de l'échelle Bark [43]). 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.*

$$z' = \frac{26.81 f}{(1960+f)} - 0.53 \quad \text{(IV.1)}$$

(IV.2)

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

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

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

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

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

$$\text{sinon} \quad z' = z \quad \text{(IV.7)}$$

(IV.8)

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

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 ([38]):

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

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 IV.3.4

IV.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:

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

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 IV.11 se récrit dans le domaine spectral sous la forme:

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

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:

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

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 IV.6 un exemple de plusieurs liftres permettant de séparer les contributions source et conduit vocal (d'après [18]).

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

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

La figure IV.7 donne un exemple de cepstre (d'après [38]).

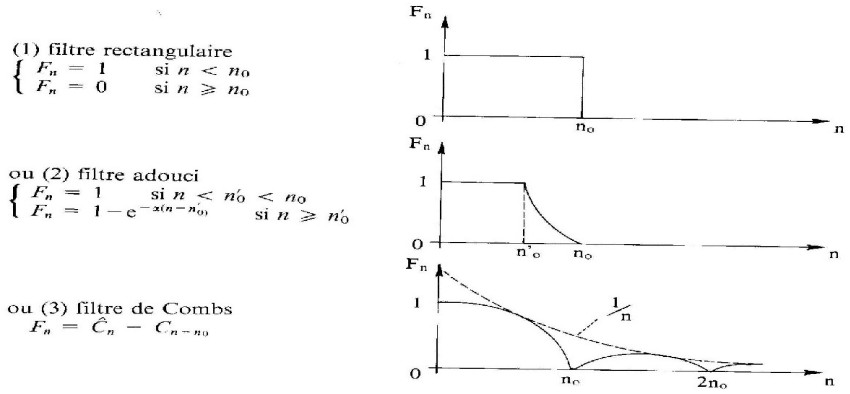


FIG. IV.6 – *Exemples de liftres (d'après [18])*

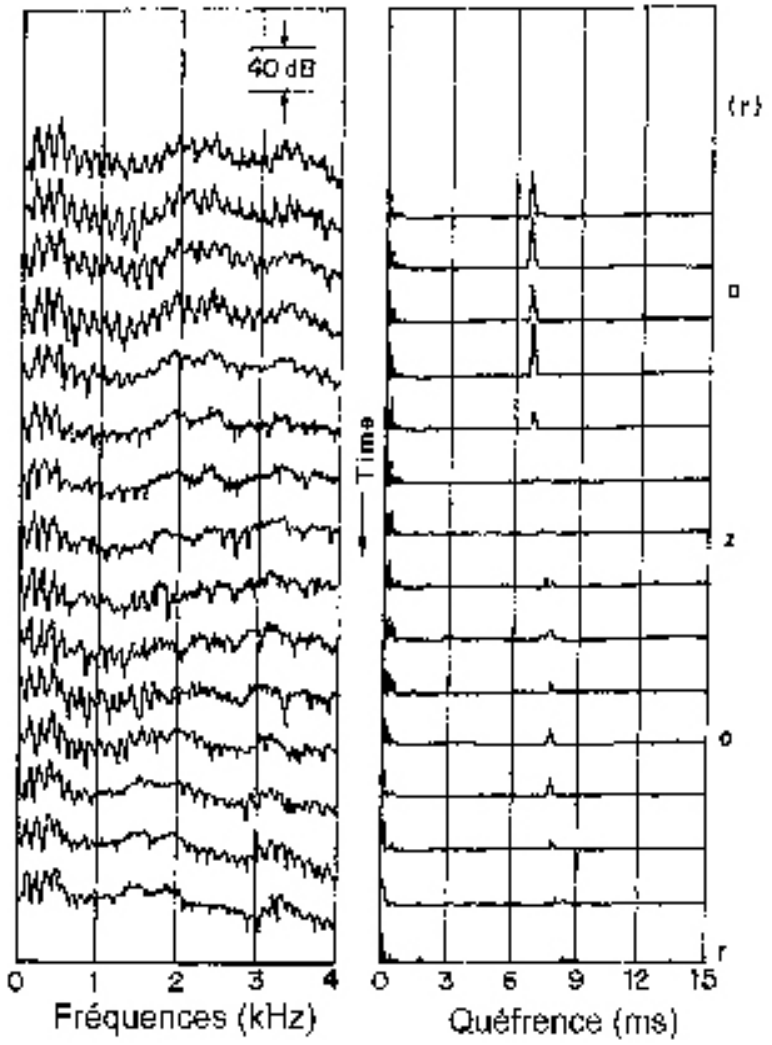


FIG. IV.7 – *Exemple de spectres à court terme (gauche) et de cepstre (droite) pour une voix d'homme prononçant le mot "razor" (d'après [38])*

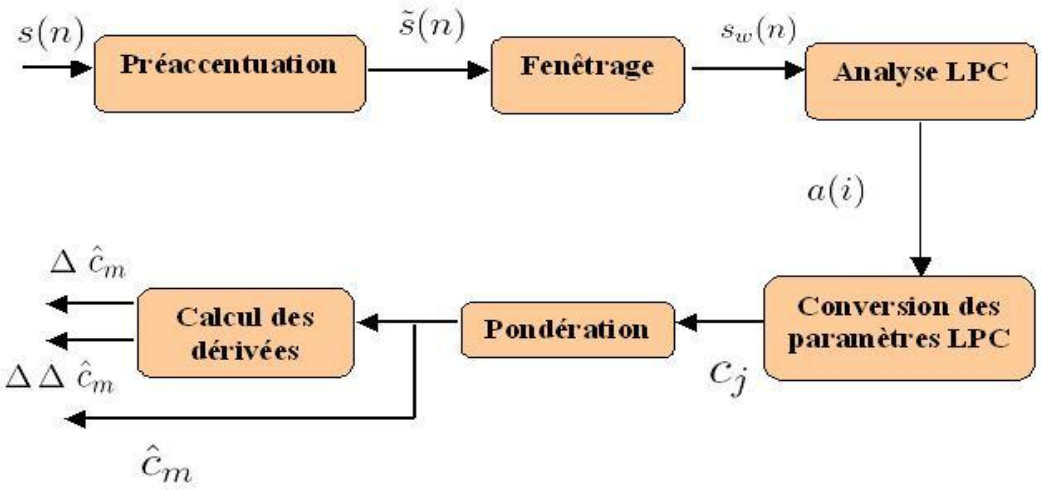


FIG. IV.8 – *Schéma bloc d'une paramétrisation LPCC*

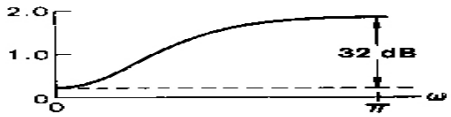


FIG. IV.9 – *Réponse en fréquence du filtre de préaccentuation (d'après [80])*

IV.3.2 La paramétrisation LPCC

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:

- 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 décrivons ci-dessous les principales étapes de la paramétrisation LPCC dont on pourra trouver un schéma en bloc sur la figure IV.8 :

1. *Préaccentuation*: Une fois numérisé, le signal $s(n)$ subit une opération de *préaccentuation*, qui consiste en un filtrage de type passe-haut qui relève le niveau des aigus. En pratique, on utilise simplement un filtre de réponse impulsionnelle finie à coefficients réels $(1, -a)$ avec $0.9 \leq a \leq 1.0$. Une valeur couramment utilisée est $a = 0.95$. Si $s(n)$ désigne le signal de parole et $\tilde{s}(n)$ le signal pré-accentué on a :

$$\tilde{s}(n) = s(n) - 0.95s(n-1) \tag{IV.15}$$

   La figure IV.9 donne la réponse en fréquence d'un tel filtre pour une valeur de $a = 0.95$ (d'après [80])

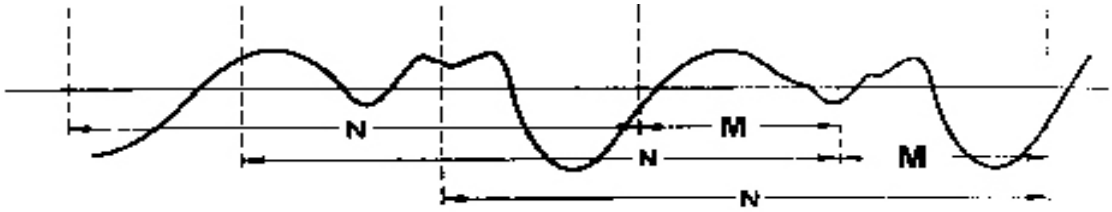


FIG. IV.10 – *Exemple de recouvrement (*overlap *en anglais) entre trames pour* $M = \frac{1}{3}N$. *(d'après [80])*

2. *Fenêtrage*: Le signal pré-accentué est alors 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 IV.10 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 :

$$s_w(n) = \tilde{s}(n)w(n) \quad \text{(IV.16)}$$

$$\text{où } w(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) \quad \text{avec } 0 \leq n \leq N-1 \quad \text{(IV.17)}$$

3. *Analyse LPC*: cette étape consiste à effectuer une modélisation AR (modélisation Autorégressive) du signal fenêtré. Dans un tel modèle on cherche à représenter le signal de parole s sous la forme:

$$s(n) = -\sum_{i=1}^{P} a(i)s(n-i) + e(n) \quad \text{(IV.18)}$$

où P est l'ordre du modèle, et $e(n)$ est l'erreur de prédiction. Il existe plusieurs méthodes pour estimer les coefficients $a(i)$ dont les méthodes basées sur l'autocorrélation et la covariance. La résolution s'appuie ensuite sur des algorithmes tels que l'algorithme de Schur ou de Levinson-Durbin dont le lecteur intéressé pourra trouver une description dans par exemple [15]. Notons cependant que les paramètres de prédiction linéaire peuvent être soit:
   - Les paramètres LPC $a(i)$ tels que décrits ci-dessus
   - Les coefficients de réflexion (encore appelées coefficients de corrélation partielle - ou PARCOR) $k(i)$
   - les coefficients LAR (Log Area ration) $g(m) = \log(\frac{1-k(m)}{1+k(m)})$

4. *Conversion des paramètres LPC en Coefficients cepstraux*: Il est ensuite possible de déduire de ces paramètres LPC, des coefficients, très utilisés en reconnaissance de parole: Les coefficients LPCC (pour Linear Predictive Cepstral Coefficients). Ils sont obtenus à partir des coefficients de prédiction linéaire en écrivant l'expansion de Laurent du filtre tout pôle $A(z)$:

$$\log(\frac{\sigma}{A(z)}) = \log\sigma + \sum_{n=1}^{\infty} c_n z^{-n} \quad \text{(IV.19)}$$

En dérivant par rapport à z^{-1}, on obtient l'expression suivante:

$$-\frac{\dot{A}(z)}{A(z)} = \sum_{n=1}^{\infty} nc_n z^{-(n-1)} \tag{IV.20}$$

En développant ensuite $A(z)$ sous la forme $A(z) = \sum_{j=0}^{p} a(j)z^{-j}$, on peut ré-écrire l'équation IV.20 sous la forme:

$$-\sum_{j=1}^{p} ja(j)z^{-(j-1)} = \sum_{n=1}^{\infty} nc_n z^{-(n-1)} \sum_{j=0}^{p} a(j)z^{-j} \tag{IV.21}$$

En égalant les termes d'égale puissance des deux termes on obtient la relation:

$$-ja(j) = \sum_{n=1}^{j} na(j-n)c_n \tag{IV.22}$$

d'où on déduit simplement l'équation IV.23 qui donne la formule de récursion pour les coefficients LPCC

$$c_j = -a(j) - \sum_{n=1}^{j-1} \frac{n}{j} a(j-n)c_n \qquad \text{pour} \quad 1 \leq j \leq p \tag{IV.23}$$

où p est l'ordre du modèle LPC et avec $c_0 = \log(\sigma^2)$

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:

$$\hat{c}_m = w(m)c_m \qquad \text{pour} \quad 1 \leq m \leq Q \tag{IV.24}$$

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:

$$w(m) = \left[1 + \frac{Q}{2} sin(\frac{\pi m}{Q})\right] \qquad \text{pour} \quad 1 \leq m \leq Q \tag{IV.25}$$

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:

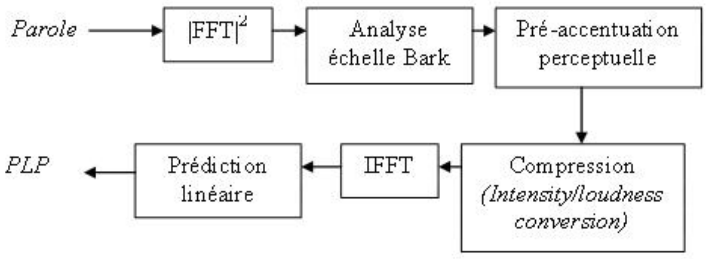


FIG. IV.11 – *Schéma bloc de l'analyse par prédiction linéaire perceptuelle (PLP) (d'après [44])*

$$\Delta c_m(t) \approx \mu \sum_{k=-K}^{K} k c_m(t+k) \tag{IV.26}$$

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

A partir des différentes étapes décrites ci-dessus, on obtient ainsi pour chaque trame t, un vecteur de paramètres acoustiques de $3Q$ composantes que l'on peut écrire:

$$\mathbf{O}_t = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_Q, \Delta c_1, \Delta c_2, \ldots \Delta c_Q, \Delta^2 c_1, \Delta^2 c_2, \ldots \Delta^2 c_Q) \tag{IV.27}$$

IV.3.3 La paramétrisation PLP

La paramétrisation PLP (*Perceptual Linear Prediction*) est également une paramétrisation populaire. On trouvera une description précise de cette paramétrisation dans [44].

Le schéma présenté sur la figure IV.11 donne les principales étapes de la paramétrisation PLP. Nous ne détaillons ci-dessous que les étapes spécifiques à cette paramétrisation.

– *Analyse par banc de filtres Bark*: le spectre $S(\Omega)$ est recalculé sur une échelle Bark en utilisant la transformation:

$$\Omega(\omega) = 6 * \log\left(\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right) \tag{IV.28}$$

Ce spectre est ensuite convolué avec les filtres caractéristiques des bandes critiques. Ce procédé est comparable à ce qui est en fait pour les coefficients cepstraux sur échelle Mel (voir ci-dessous) sauf que la forme des filtres est ici différente et est donnée en échelle Bark par:

$$\Psi(\Omega) = \begin{cases} 0 & \text{pour } \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & \text{pour } -1.3 \leq \Omega \leq -0.5 \\ 1 & \text{pour } -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & \text{pour } 0.5 \leq \Omega \leq 2.5 \\ 0 & \text{pour } \Omega > 2.5 \end{cases} \tag{IV.29}$$

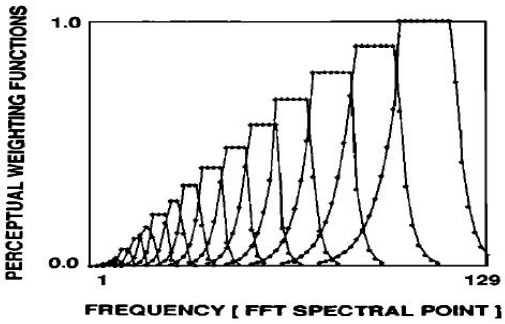


FIG. IV.12 – *Banc de filtres en bandes critiques (d'après [44])*

La convolution entre le spectre $S(\omega)$ et le filtre de gain $\Psi(\Omega)$ donne les échantillons du spectre de puissance sur une échelle Bark:

$$\theta(\Omega_i) = \sum_{\omega=-1.3}^{2.5} S(\Omega - \Omega_i)\Psi(\Omega) \tag{IV.30}$$

Le banc de filtres correspondant est donné sur la figure IV.12

– *Préaccentuation perceptuelle* : cette étape consiste à prendre en compte les variations de sensibilité de l'oreille avec la fréquence. Elle est réalisée en pré-accentuant le spectre de puissance $\Theta(\Omega(\omega))$ précédemment calculé à l'aide de la fonction $E(\omega)$ qui simule la sensibilité de l'oreille à -40 dB : $\Xi(\Omega(\omega)) = E(\omega)\Theta(\Omega(\omega))$. Dans l'implémentation originale, la fonction $E(\omega)$ est approximée par :

$$E(\omega) = \frac{(\omega^2 + 56.8 \times 10^6)\omega^4}{(\omega^2 + 6.3 \times 10^6)^2(\omega^2 + 0.38 \times 10^9)} \tag{IV.31}$$

qui est valable pour les fréquences inférieures à 5000 Hz. Pour des fréquences supérieures un terme correctif est rajouté.

– *Compression (Intensity-loudness conversion)*: C'est la dernière opération avant le calcul des coefficients de prédiction linéaire et elle correspond à une compression en racine cubique sous la forme: $\phi(\Omega) = \Xi(\Omega)^{0.33}$. C'est ici une approximation de la loi de Stevens et elle simule la relation non-linéaire entre l'intensité d'un son et la sensation de puissance sonore correspondante.

On peut ensuite en déduire des coefficients cepstraux PLP en suivant la même approche que pour les coefficients LPCC.

IV.3.4 La paramétrisation MFCC

La paramétrisation MFCC (Mel-Frequency Cepstral coefficients) est probablement la paramétrisation la plus répandue dans les systèmes de reconnaissance actuels.

De même que pour les coefficients LPCC, un certain nombre d'étapes sont nécessaires pour cette paramétrisation. Nous ne développerons ci-dessous que les étapes qui ne se retrouvent pas dans la paramétrisation LPCC:

– *Fenêtrage du signal* similairement à la paramétrisation LPCC

– *Calcul de la transformée de Fourier rapide* (FFT)pour chaque trame du signal de parole

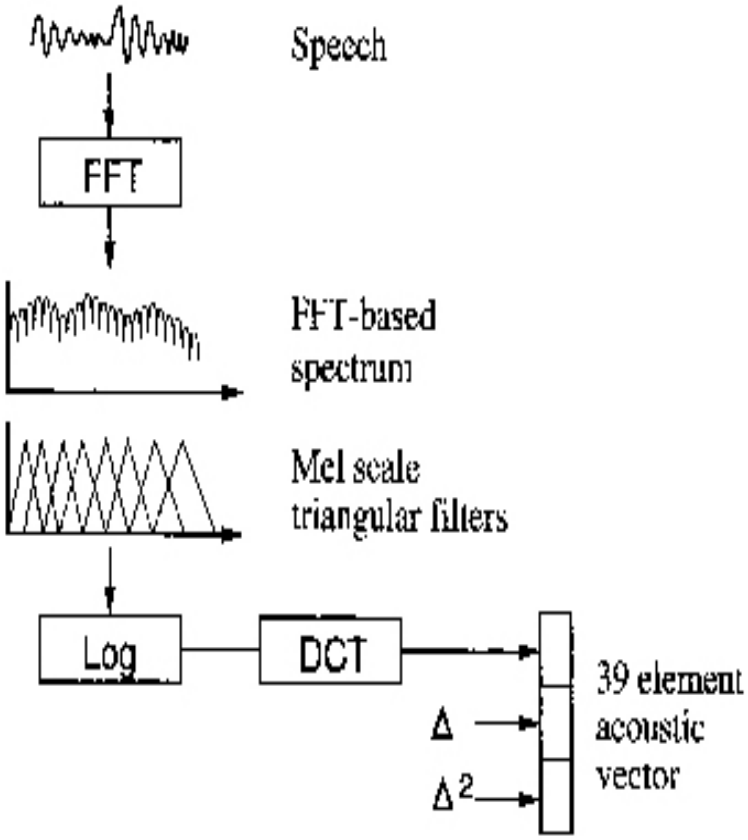


FIG. IV.13 – *Schéma bloc de la paramétrisation MFCC (*MFCC-based Front-End processor *d'après [38])*

- *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).
- *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 $\tilde{S}(k)$:

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n(k - \frac{1}{2})\frac{\pi}{K}\right] \qquad \text{pour} \quad n = 1,2,\ldots,L \tag{IV.32}$$

où L est le nombre de coefficients cepstraux désirés.

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 IV.13 donne un schéma bloc de cette implémentation classique.

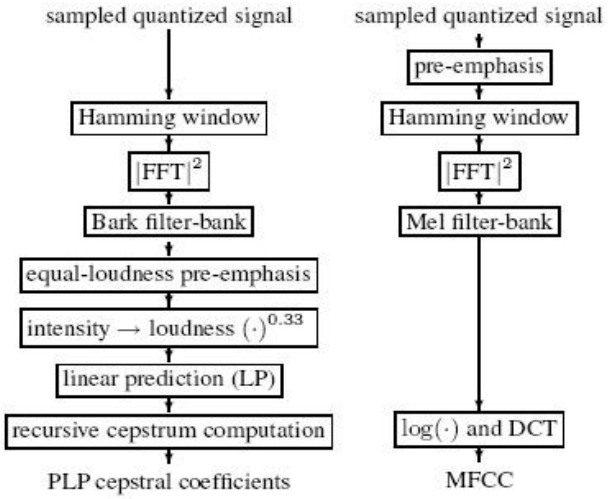


FIG. IV.14 – *Principales étapes pour le calcul des PLP (à gauche) et des MFCC (à droite) d'après [48])*

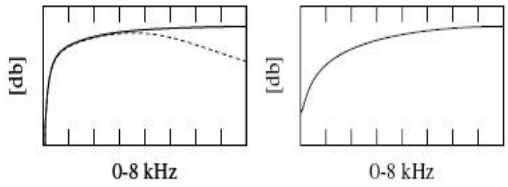


FIG. IV.15 – *Préaccentuation perceptuelle pour les PLP (à gauche), la courbe en pointillé intégrant un facteur correctif pour les fréquences au dessus de 5 kHz; La courbe de préaccentuation traditionnelle pour les MFCC est donnée à droite (d'après [48])*

IV.3.5 Comparaison entre les PLP et les MFCC

La figure IV.14 représente les principales étapes des paramétrisations MFCC et coefficients cepstraux PLP, ce qui permet de mieux comprendre les principales différences entre ces deux paramétrisations. Si en apparence, les deux paramétrisations semblent très différentes, une étude plus précise permet de voir qu'elles partagent de nombreux points communs.

Par exemple, la pré-accentuation apparait très semblable dans les deux approches surtout lorsque le terme correctif de la courbe de préaccentuation perceptuelle pour les paramètres PLP n'est pas utilisé (voir figure IV.15).

Il est aussi important de remarquer que dans les deux approches, on retrouve une estimation de l'enveloppe spectrale par lissage du spectre (en utilisant un modèle tout pôle pour les PLP et en tronquant les coefficients cepstraux pour les MFCC).

Par contre, si les deux utilisent des bancs de filtres s'inspirant de la perception, la forme et la largeur de bande de ces filtres diffèrent de manière non négligeable (voir figure IV.16).

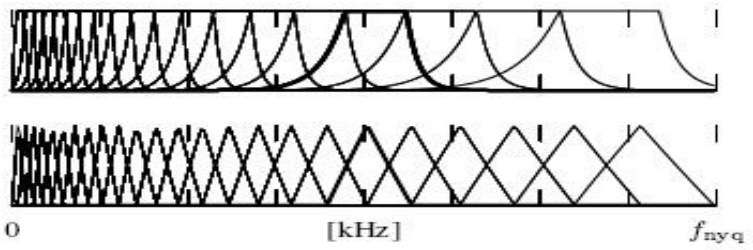


FIG. IV.16 – *Banc de filtres Bark (en haut) et Mel (en bas) (d'après [48])*

IV.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.

IV.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.

IV.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:

$$d(S,S')^p = \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^p \frac{d\omega}{2\pi} \tag{IV.33}$$

La figure IV.17 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:

$$d_{lpc}(S,S')^p = \int_{-\pi}^{\pi} |\log \frac{\sigma^2}{|A(e^{j\omega})|^2} - \log \frac{\sigma'^2}{|A'(e^{j\omega})|^2}|^p \frac{d\omega}{2\pi} \tag{IV.34}$$

La figure IV.18 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 IV.17 que la différence est beaucoup plus régulière et est, bien sur, moins sensible aux différences de fréquence fondamentale.

IV.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:

$$d_2^2 = \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 \frac{d\omega}{2\pi} \tag{IV.35}$$

$$= \sum_{n=-\infty}^{\infty} (c_n - c_n^{'})^2 \tag{IV.36}$$

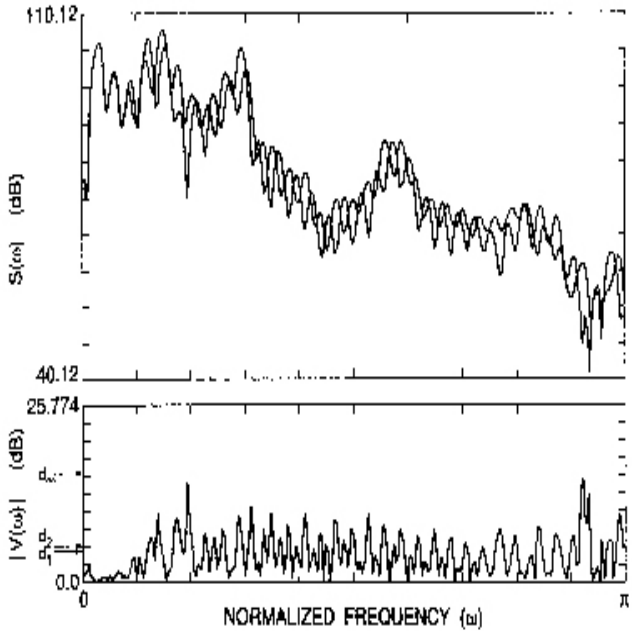


FIG. IV.17 – *Spectre d'amplitude $S(\omega)$ et $S'(\omega)$ (en haut) et le module de leur différence logarithmique (en bas) (d'après [80])*

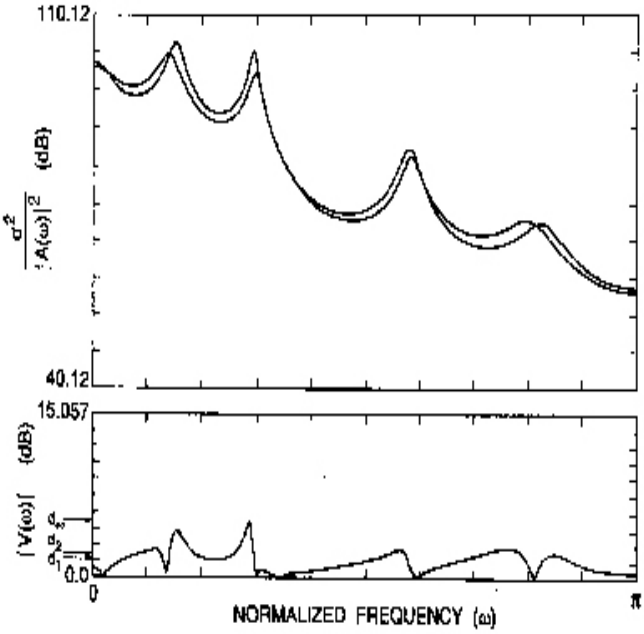


FIG. IV.18 – *Modèles LPC $\frac{\sigma^2}{|A(e^{j\omega})|^2}$ et $\frac{\sigma'^2}{|A'(e^{j\omega})|^2}$ (en haut) et le module de leur différence logarithmique (en bas) (d'après [80])*

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):

$$d_2^2 = \sum_{n=1}^{L} (c_n - c_n^{'})^2 \tag{IV.37}$$

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 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 IV.25. Dans ce cas, la distance s'écrira:

$$d_w^2 = \sum_{n=1}^{L} (w(n)c_n - w(n)c_n^{'})^2 \tag{IV.38}$$

IV.4.4 Mesures de distorsion et rapport de vraisemblance

La différence Log-spectrale est à l'origine d'un grand nombre de mesures de distorsion. L'une de ces mesures, proposée par Itakura-Saito (qu'on appelera souvent la distance d'Itakura-Saïto dans la littérature) est donnée par:

$$d_{IS}(S,S') = \int_{-\pi}^{\pi} \frac{S(\omega)}{S'(\omega)} \frac{d\omega}{2\pi} - \log \frac{\sigma_\infty^2}{\sigma_\infty^{'2}} - 1 \tag{IV.39}$$

où σ_∞^2 et $\sigma_\infty^{'2}$ sont les erreurs de prédictions définies telles que (voir [80],p154-155, pour plus de précisions):

$$\sigma_\infty^2 = \exp\left(\int_{-\pi}^{\pi} \log S(\omega) \frac{d\omega}{2\pi} \right) \tag{IV.40}$$

La distance d'Itakura-Saïto possède un certain nombre de propriétés et il est notamment intéressant de remarquer que cette distance n'en est pas vraiment une puisqu'elle n'est pas symmétrique.

On peut également dériver d'autres distances couramment utilisées à partir de la distance d'itakura-Saïto. On nommera notamment la distance d'Itakura qui s'exprime sous la forme:

$$d_I = d_{IS}(\sigma_p^2/|A_p|^2, \sigma^2/|A|^2) \tag{IV.41}$$

où $\sigma_p^2/|A_p|^2$ est le modèle optimal pour $S(\omega)$ et $\sigma^2/|A|^2$ est le modèle de $S'(\omega)$.

On peut également définir la distorsion appelée "rapport de vraisemblance" qui s'écrit:

$$\begin{aligned} d_{LR}(\frac{1}{|A_p|^2}, \frac{1}{|A|^2}) &= d_{IS}(\frac{1}{|A_p|^2}, \frac{1}{|A|^2}) && \text{(IV.42)} \\ &= \int_{-\pi}^{\pi} \frac{|A(ej\omega)|^2}{|A_p(ej\omega)|^2} \frac{d\omega}{2\pi} - 1 && \text{(IV.43)} \\ &= \frac{\mathbf{a}\mathbf{R_p}\mathbf{a}}{\sigma_p^2} - 1 && \text{(IV.44)} \end{aligned}$$

où **a**, R_p et σ_p sont respectivement les coefficients du modèle prédictif de $S'(\omega)$, la matrice d'autocorrélation et l'erreur résiduelle du modèle de $S(\omega)$

IV.4.5 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 $d^2_{2\Delta}$ qui est proche de la distance spectrale différentielle:

$$d^2_{2\Delta} = \int_{-\pi}^{\pi} |\frac{\partial \log S(\omega,t)}{\partial t} - \frac{\partial \log S'(\omega,t)}{\partial t}|^2 \frac{d\omega}{2\pi} \quad \text{(IV.45)}$$

$$\simeq \sum_{n=-\infty}^{\infty} (\Delta c_n - \Delta c_n^{'})^2 \quad \text{(IV.46)}$$

De même, pour les dérivées secondes on aura:

$$d^2_{2\Delta(2)} = \int_{-\pi}^{\pi} |\frac{\partial^2 \log S(\omega,t)}{\partial^2 t} - \frac{\partial^2 \log S'(\omega,t)}{\partial^2 t}|^2 \frac{d\omega}{2\pi} \quad \text{(IV.47)}$$

$$\simeq \sum_{n=-\infty}^{\infty} (\Delta c_n^{(2)} - \Delta c_n^{'(2)})^2 \quad \text{(IV.48)}$$

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

$$d_{2,\Delta,\Delta^{(2)}} = \gamma_1 d_2^2 + \gamma_2 d^2_{2\Delta} + \gamma_3 d^2_{2\Delta^{(2)}} \quad \text{(IV.49)}$$

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

IV.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:

$$d(\chi,\xi) = \sum_{i_x=1}^{T_x} d(i_x,i_y) \quad \text{(IV.50)}$$

où i_x et i_y vérifient la relation

$$i_y = \frac{T_y}{T_x} i_x \quad \text{(IV.51)}$$

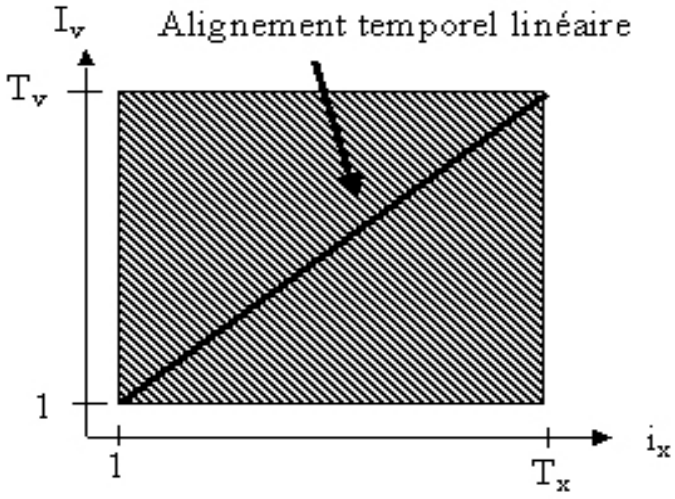


FIG. IV.19 – *Alignement temporel linéaire*

Une illustration de cet alignement temporel linéaire est donnée figure IV.19

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:

$$i_x = \phi_x(k) \text{ pour } k = 1,2,\ldots,T \quad \text{(IV.52)}$$

$$i_y = \phi_y(k) \text{ pour } k = 1,2,\ldots,T \quad \text{(IV.53)}$$

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

$$d_\phi(\chi,\xi) = \sum_{k=1}^{T} d(\phi_x(k),\phi_y(k)))m(k)/M_\phi \quad \text{(IV.54)}$$

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 IV.20 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:

$$d(\chi,\xi)) = \min_\phi d_\phi(\chi,\xi) \quad \text{(IV.55)}$$

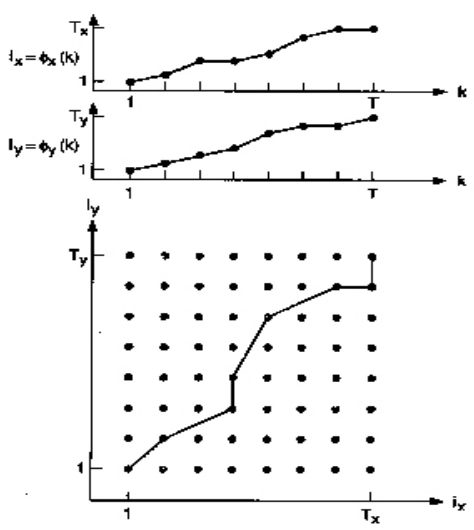


FIG. IV.20 – *Exemple d'alignement dynamique (d'après [80]). La ligne en trait plein indique le chemin le long duquel la distance $d(\phi_x(k),\phi_y(k)))$ est évaluée.*

IV.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 ([11]).

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) 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:

$$D(i_x,i_y) = \min_{\phi_x,\phi_y,T'} \sum_{k=1}^{T'} d(\phi_x(k),\phi_y(k))m(k) \tag{IV.56}$$

où

$$\phi_x(T') = i_x \ ; \ \phi_y(T') = i_y \tag{IV.57}$$

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:

$$M_\phi = \sum_{k=1}^{T} m(k) \tag{IV.58}$$

L'algorithme de programmation dynamique avec contraintes devient alors:

$$D(i_x,i_y) = \min_{(i'_x,i'_y)} [D(i'_x,i'_y) + \zeta((i'_x,i'_y),(i_x,i_y))] \tag{IV.59}$$

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

$$\zeta((i'_x,i'_y),(i_x,i_y)) = \sum_{l=0}^{L_s} d(\phi_x(T'-l),\phi_y(T'-l)m(T'-l) \tag{IV.60}$$

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 :

$$\phi_x(T'-Ls) = i'_x \text{ et } \phi_y(T'-Ls) = i'_y \tag{IV.61}$$

La figure IV.21 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 IV.21)

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 IV.22)
- 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 IV.21 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 (IV.59)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 (IV.59) 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.

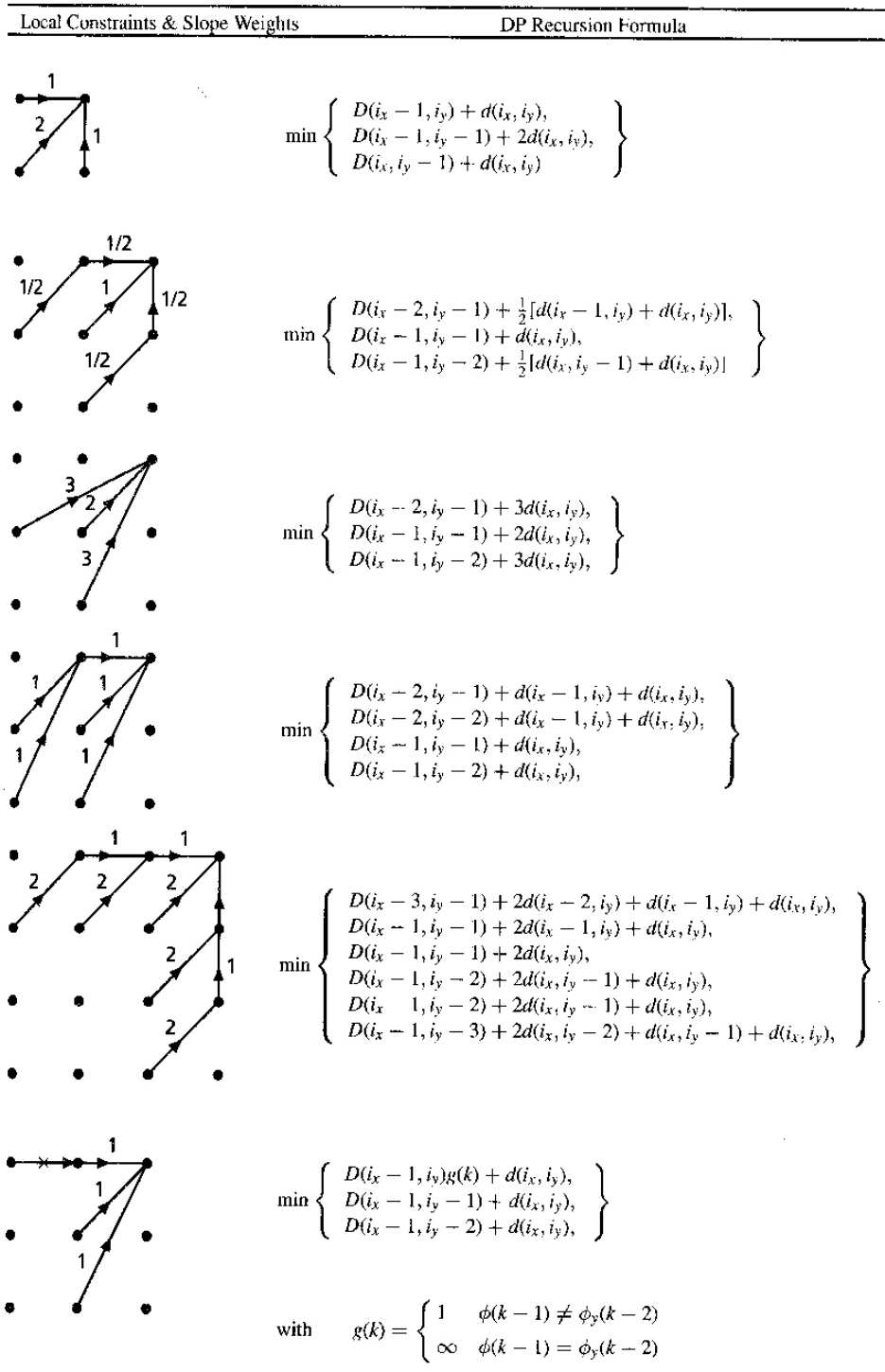


FIG. IV.21 – *Contraintes locales et pondération (d'après [80]).*

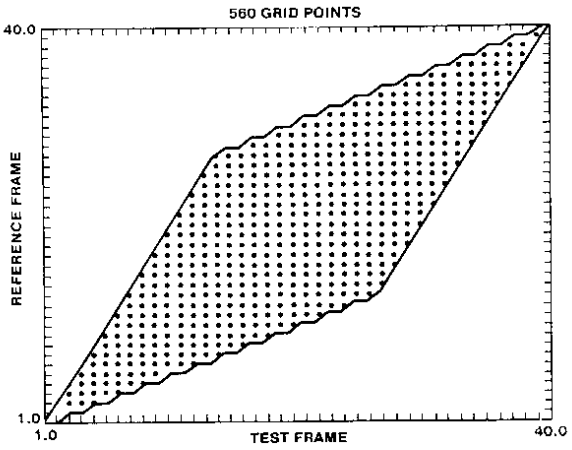


FIG. IV.22 – *Région de recherche (contraintes temporelles permettant un taux de compression/expansion local de 2:1)*

IV.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) de référence. Plusieurs approches ont été proposées pour adapter l'algorithme de programmation dynamique (voir [15],[80]).

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 IV.23 donne un exemple de chemin DTW dans le cas de mots enchaînés.

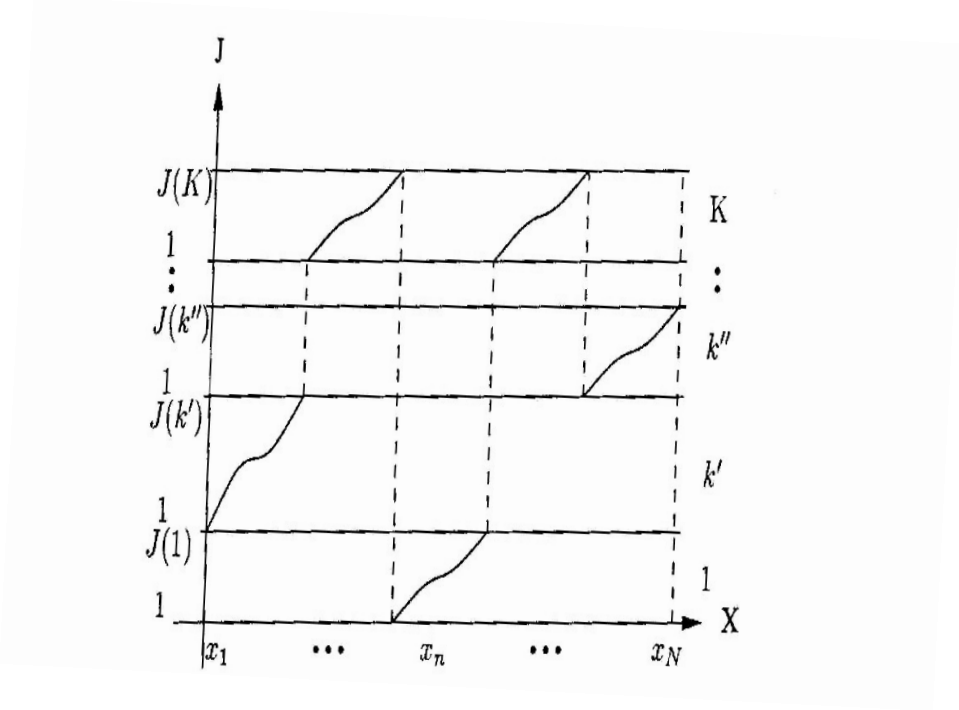


FIG. IV.23 – *Exemple de chemin DTW dans le cas de mots enchaînés. Dans cet exemple la phrase prononcée contenait la séquence de mots k' - K - 1 - K - k" (d'après [15]).*

IV.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.

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 ([16]). 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 ([15]). 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.

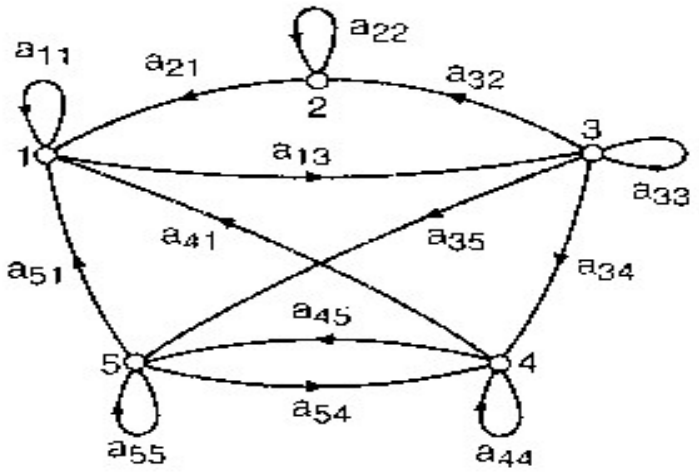


FIG. IV.24 – *Modèle de markov à 5 états avec transitions entre états (d'après [80]).*

IV.6 Les modèles de Markov cachés (HMM)

Notons que cette partie est très largement inspirée de [80]) que le lecteur intéressé par plus de détails pourra consulter.

Précisons également que ce cours ne propose qu'une introduction aux modèles de Markov puisqu'ils seront vus plus en détail dans le cadre d'un module (brique) plus avancé.

Dans cette partie, nous présentons ainsi une approche statistique très largement utilisée en reconnaissance vocale : l'approche par modèles de Markov cachés (ou HMM pour *Hidden Markov Models*). L'hypothèse sous-jacente des modèles de Markov cachés (ou de n'importe quel modèle statistique) est que le signal de parole peut être bien représenté comme un processus aléatoire paramétrique et que les paramètres de ce processus stochastique peuvent être déterminés (ou estimés) d'une façon précise et bien définie.

La théorie des modèles de Markov cachés est ancienne et fut appliquée au traitement de la parole dès le milieu des années 1970 ([8], [7]).

IV.6.1 Chaînes de Markov discrètes

En guise d'introduction aux chaînes de Markov, nous allons considérer dans un premier temps le système représenté figure IV.24 qui est un modèle à $N = 5$ états. A chaque instant, ce système est dans l'un des N états, et à des instants régulièrement espacés (on est ici à temps discret), le système change d'état (ou reste dans le même état) en fonction d'un ensemble de probabilités associés à l'état courant.

A partir des notations suivantes:

$$t \quad = 1,2... \quad \text{sont les instants de changement d'état} \tag{IV.62}$$

$$q_t \quad \text{est l'état à l'instant } t \tag{IV.63}$$

nous notons la probabilité d'être dans l'état j sachant que l'on a été dans l'état i au temps $t-1$ et dans l'état k à l'état $t-2$, etc ... sous la forme:

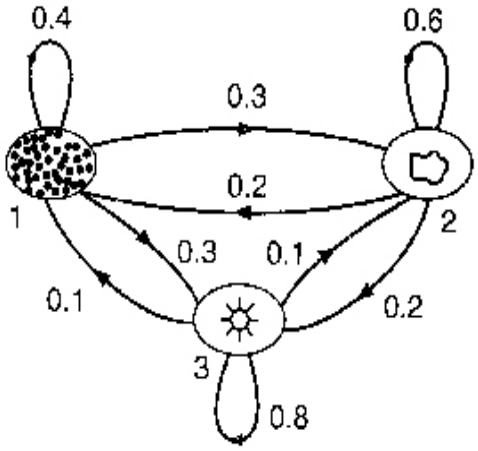


FIG. IV.25 – *Modèle météorologique à trois états (d'après [80]).*

$$P[q_t = j|q_{t-1} = i,q_{t-2} = k,\ldots] \tag{IV.64}$$

Pour le cas particulier de chaînes de Markov du premier ordre, la probabilité est "tronquée" à la connaissance de l'état précédent, soit:

$$P[q_t = j|q_{t-1} = i,q_{t-2} = k,\ldots] = P[q_t = j|q_{t-1} = i] \tag{IV.65}$$

De plus, nous ne considérons ici que les processus dont le terme de droite de l'équation IV.64 est indépendant du temps, ce qui amène à considérer un ensemble de probabilités de changement d'état (indépendantes du temps) que l'on notera a_{ij}:

$$a_{ij} = P[q_t = j|q_{t-1} = i] \qquad \text{pour} \quad 1 \leq i,j \leq N \tag{IV.66}$$

avec les propriétés suivantes:

$$a_{ij} \geq 0 \quad \forall j,i \tag{IV.67}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i \tag{IV.68}$$

Pour mieux comprendre le principe général, considérons le modèle de Markov à trois états donné figure IV.25 comme modèle météorologique. Nous supposons que chaque jour le temps est observé à une heure précise (par exemple tous les jours à midi).

Nous définissons alors les trois états correspondants à trois situations météorologiques différentes:

- *L'état 1:* Précipitations (pluie, neige ou grêle)
- *L'état 2:* Nuageux
- *L'état 3:* Beau temps

On suppose, de plus, que le temps au jour t ne peut être décrit que par l'un des états ci-dessus. On définit la matrice A de transition entre états:

$$A = \begin{pmatrix} 0.4_{11} & 0.3_{12} & 0.3_{13} \\ 0.2_{21} & 0.6_{22} & 0.2_{23} \\ 0.1_{31} & 0.1_{32} & 0.8_{33} \end{pmatrix} \tag{IV.69}$$

Étant donné ce modèle (voir figure IV.25), il est possible de calculer la probabilité que le temps pour les 8 prochains jours soit: Soleil, Soleil, Soleil, Pluie, Pluie, Soleil, Nuageux, Soleil.

Pour cela, on définit la séquence d'observations **O** sous la forme:

$$\begin{array}{lccccccccc} \mathbf{O} & = & (& \text{Soleil}, & \text{Soleil}, & \text{Soleil}, & \text{Pluie}, & \text{Pluie}, & \text{Soleil}, & \text{Nuageux}, & \text{Soleil}, &) \\ & = & (& 3, & 3, & 3, & 1, & 1, & 3, & 2, & 3, &) \\ \text{Jour} & & & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & \end{array} \tag{IV.70}$$

et on veut calculer $P(\mathbf{O}|Model)$, la probabilité d'observer la séquence d'observations **O**, étant donné le modèle de la figure IV.25. On peut alors déterminer directement $P(\mathbf{O}|Model)$ comme:

$$\begin{aligned} P(\mathbf{O}|Model) &= P[3,3,3,1,1,3,2,3,|Model] && \text{(IV.71)} \\ &= P[3]P[3|3]^2P[1|3]P[1|1]P[3|1]P[2|3]P[3|2] && \text{(IV.72)} \\ &= \pi_3 \cdot (a_{33})^2 a_{31}a_{11}a_{13}a_{32}a_{23} && \text{(IV.73)} \\ &= (1.0)(0.8)^2(0.1)(0.4)(0.3)(0.1)(0.2) && \text{(IV.74)} \\ &= 1.536 \times 10^{-4} && \text{(IV.75)} \end{aligned}$$

où nous avons utilisé la notation suivante pour la probabilité π_i de l'état initial i :

$$\pi_i = P[q_1 = i], \qquad 1 \leq i \leq N \tag{IV.76}$$

IV.6.2 Extensions aux modèles de Markov cachés

Nous avons considéré jusqu'à maintenant que chaque état correspondait à un événement déterministe et observable. Ainsi, la sortie de ces sources dans chaque état n'est pas aléatoire. Ce modèle est trop restrictif et, nous allons ainsi présenté ci-dessous une extension du concept des modèles Markoviens pour inclure la notion où l'observation est une fonction probabiliste de l'état, c'est à dire que le modèle résultant (appelé modèle de Markov Caché) est un processus doublement stochastique avec un processus stochastique sous-jacent qui n'est pas directement observable (il est caché) mais qui ne peut être observé qu'à travers un autre processus stochastique qui produit la séquence d'observation.

Pour illustrer ces concepts de base, nous allons prendre deux exemples simples: le modèle du jet de pièce (pile ou face) et le désormais célèbre exemple des boules et des urnes.

Modèle Pile ou Face

L'expérience suivante est menée. Un nombre T de tirage est effectué en utilisant une ou plusieurs pièces. Ce tirage est effectué dans une pièce séparée et seul le résultat nous est communiqué, de telle sorte qu'on ne sait pas quelle pièce à été utilisée à tel moment. On obtient après T tirages une séquence d'expériences cachées produisant une séquence d'observation qui est par exemple:

$$\begin{array}{lcccccccc} \mathbf{O} & = & (& \mathbf{O_1} & \mathbf{O_2} & \mathbf{O_3} & \dots & \mathbf{O_T} &) \\ & = & (& \text{F} & \text{P} & \text{F} & \dots & \text{F} & \end{array}$$

où P et F représentent respectivement un tirage "Pile" et "Face".

Il y a maintenant plusieurs façon de construire un modèle HMM (pour *Hidden Markov Model*). Le premier modèle qui vient à l'esprit est le modèle *a)* donné sur la figure IV.26. Nous avons ici un modèle observable à 2 états et où le seul paramètre à connaître est la probabilité de l'occurrence du tirage "Face". Un second modèle que l'on peut utiliser pour expliquer la séquence observée est le modèle *b)*donné figure IV.26. Dans ce modèle, nous avons 2 états où chacun correspond à une pièce différente (qui possède ainsi un biais différent). Ainsi, chaque état est caractérisé par une probabilité de distribution entre "Pile et "face" et les transitions entre états sont caractérisées par une matrice de transition. On peut alors extrapoler ce modèle à un nombre de pièces plus grand. Le troisième modèle (le modèle *c)*) de la figure IV.26 donne un exemple lorsque 3 pièces sont utilisées pour l'expérience.

Sachant que plusieurs modèles sont donc possibles, il est alors naturel d'essayer de choisir le meilleur modèle pour décrire les observations. A priori, on pourrait penser que plus le modèle est complexe meilleur sera le modèle. En pratique, cela n'est pas nécessairement vrai puisque la taille des modèles doit rester limitée, qu'il est important de bien pouvoir apprendre le modèle et qu'il est nécessaire de savoir si les données sont suffisamment importantes pour correctement apprendre un modèle complexe. Nous pouvons cependant faire ici une remarque. Si on sait qu'une seule pièce de monnaie a été utilisée pour ces expériences, alors clairement le modèle du bas de la figure IV.26 (modèle (c)) n'est pas approprié car alors les données d'apprentissage ne permettront pas de correctement apprendre le modèle.

Modèle des boules et des urnes

Le système des boules et des urnes permet de présenter une situation légèrement plus complexe. Nous supposons que nous avons N urnes placées dans une pièce et que chaque urne contient une grande quantité de boules colorées (voir figure IV.27).

Le nombre de couleurs différentes est M. Le tirage d'une séquence d'observations se fait de la façon suivante:

- Une urne est sélectionnée selon une procédure aléatoire
- Une boule est ensuite tirée de l'urne préalablement sélectionnée et la couleur de cette boule constitue l'observation
- La boule est replacée dans l'urne et une nouvelle urne peut alors être sélectionnée pour le tirage suivant

En itérant ce processus T fois on obtient une séquence d'observations sous la forme:

$$\mathbf{O} = (\text{bleu, rouge, jaune, orange, violet, } \dots, \text{ rouge}) \tag{IV.77}$$

Notons que le tirage est effectué dans une autre pièce ce qui ne permet pas de savoir dans quelle urne a été tirée chaque boule. En effet, comme chaque urne peut contenir des boules de toutes les couleurs, la connaissance de la suite d'observations ne permet pas de déduire de façon immédiate de quelle urne a été tirée chaque boule.

Comme pour l'expérience du "pile ou face", nous pouvons prendre ici un modèle de Markov assez simple qui peut décrire cette expérience. Ce modèle possédera autant d'états que d'urnes pour lesquels des probabilités de couleurs sont définis. Le choix des urnes dans lesquelles les

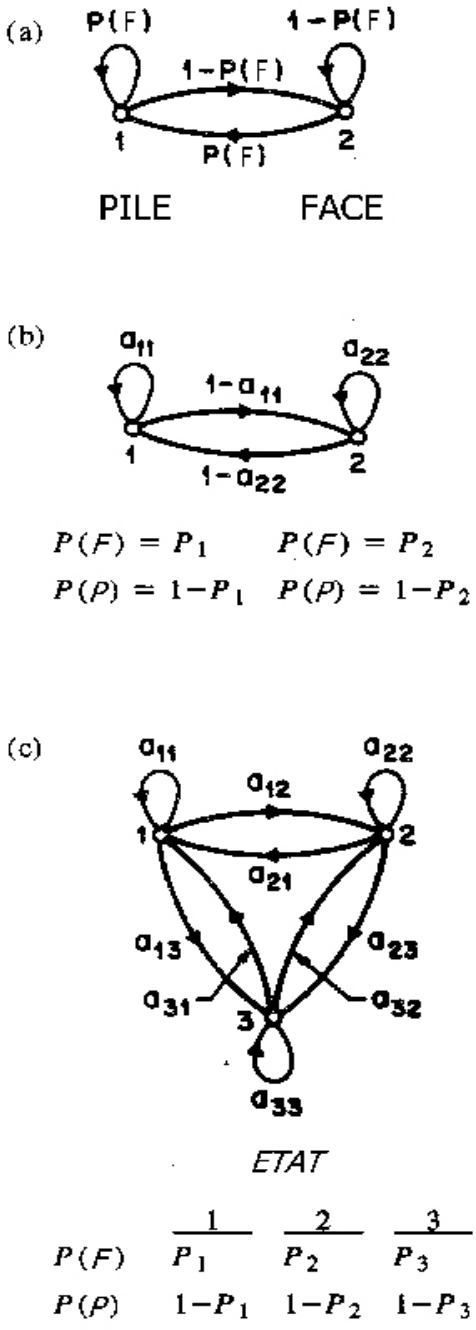


FIG. IV.26 – *Trois modèles possibles pour le tirage "pile ou face" d'une ou plusieurs pièces de monnaie (d'après [80]).*

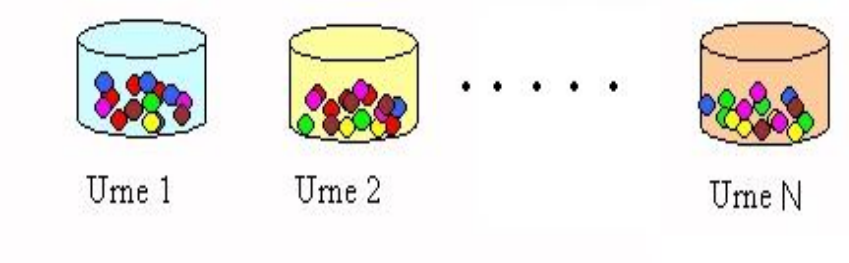


FIG. IV.27 – *Modèle à N états des urnes et boules (d'après [80]).*

boules sont tirées est alors décrit par une matrice de transitions entre états (équivalent à la matrice A de l'exemple précédent).

Caractérisation des HMM

Nous avons pu nous familiariser avec les HMM à partir des 2 exemples simples précédents. Nous pouvons maintenant donner de manière un peu plus formelle les paramètres d'un HMM. Ainsi, un modèle HMM pour des données d'observations discrètes telles que celles des boules et et des urnes est caractérisé par :

- *Le nombre N d'états* du modèle: même si les états sont cachés, il est clair (comme on a pu le voir dans les deux exemples simples) que l'on peut souvent rattacher un phénomène physique à l'ensemble des états du modèle. Lorsque tous les états sont interconnectés entre eux, on parlera de *modèle ergodique*. Lorsqu'on n'autorisera qu'une progression de gauche à droite on parlera de *modèle HMM gauche-droite*. Les différents états sont notés 1,2,...N et l'état au temps t est noté q_t
- Le nombre M de symboles distincts d'observation par état (soit la taille de l'alphabet). Ces symboles correspondent à"Pile" et "Face" dans l'expérience du tirage "pile ou face". Nous noterons ces symboles sous la forme $V = \mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_M$.
- la matrice $A = a_{ij}$ de transition entre états où

$$a_{ij} = P[q_t = j|q_{t-1} = i] \qquad \text{pour} \quad 1 \leq i,j \leq N \tag{IV.78}$$

  Notons que si toutes les transitions sont autorisées alors tous les a_{ij} seront strictement positifs. Dans les autres cas, un certain nombre de transitions seront nulles.
- La distribution de probabilité $B = b_j(k)$ d'observation des symboles pour laquelle

$$b_j(k) = P[\mathbf{o}_t = \mathbf{v}_k|q_t = j], \qquad \text{pour} \quad 1 \leq k \leq M \tag{IV.79}$$

  définit la distribution de probabilité des symboles dans l'état j,$j = 1,2,\ldots N$.
- la distribution $\pi = \pi_j$ de la distribution de l'état initial pour laquelle:

$$\pi_j = P[q_1 = j], \qquad \text{pour} \quad 1 \leq j \leq N \tag{IV.80}$$

Ainsi, on peut résumer en disant que la spécification complète d'un HMM inclut:

- La spécification des deux paramètres N et M du modèle
- la spécification des symboles d'observation

– et la spécification des probabilités A, B, et π.

Par convention, on notera $\lambda = (A,B,\pi)$ pour désigner le modèle complet. Ce modèle inclut une mesure de probabilité pour $\mathbf{O}$, soit $P(\mathbf{O}|\lambda)$.

Les trois problèmes des HMM

Rabiner et Juang ([80]) ont résumé l'utilisation des HMM à trois principaux problèmes. Ces problèmes peuvent être énoncés comme suit:

– *Problème 1: Évaluer la probabilité d'une séquence d'observations.* Ou encore, connaissant la séquence d'observation $\mathbf{O} = (\mathbf{o_1},\mathbf{o_2},\mathbf{o_3},\ldots,\mathbf{o_T})$ et le modèle $\lambda = (A,B,\pi)$, comment peut-on calculer $P(\mathbf{O}|\lambda)$ qui est la probabilité de la séquence d'observations, connaissant le modèle ,
– *Problème 2: Retrouver la séquence d'états optimale.* Ou encore, connaissant la séquence d'observation $\mathbf{O} = (\mathbf{o_1},\mathbf{o_2},\mathbf{o_3},\ldots,\mathbf{o_T})$ et le modèle $\lambda = (A,B,\pi)$, comment choisit-on la séquence d'état $q = (q_1,q_2,q_3,\ldots,q_T)$ qui est optimale au sens d'un certain critère (i.e. la séquence d'état qui "explique" au mieux les observations)?
– *Problème 3: Ré-estimer les paramètres du modèle.* Comment ajuste-t-on les paramètres du modèle $\lambda = (A,B,\pi)$ pour maximiser $P(\mathbf{O}|\lambda)$ qui est la probabilité de la séquence d'observation connaissant le modèle.

Problème 1: Évaluer la probabilité d'une séquence d'observations Nous souhaitons ici, évaluer la probabilité $P(\mathbf{O}|\lambda)$ de la séquence d'observations $\mathbf{O} = (\mathbf{o_1},\mathbf{o_2},\mathbf{o_3},\ldots,\mathbf{o_T})$, connaissant le modèle $\lambda = (A,B,\pi)$. La façon la plus intuitive de faire serait d'énumérer toutes les séquences d'états de taille T. Nous aurons alors N^T séquences possibles. Soit, $\mathbf{q}$ l'une de ces séquences d'états. On notera:

$$\mathbf{q} = (q_1,q_2,\ldots q_T) \tag{IV.81}$$

où q_1 est l'état initial. En supposant que les observations sont statistiquement indépendantes, la probabilité d'observer la séquence $\mathbf{O}$ étant donné la séquence d'états $\mathbf{q}$ est donnée par:

$$P(\mathbf{O}|\mathbf{q},\lambda) = \prod_{t=1}^{T} P(\mathbf{o}_t|q_t,\lambda) \tag{IV.82}$$

On en déduit alors:

$$P(\mathbf{O}|\mathbf{q},\lambda) = b_{q_1}(\mathbf{o}_1)\cdot b_{q_2}(\mathbf{o}_2)\ldots b_{q_T}(\mathbf{o}_T) \tag{IV.83}$$

On peut également écrire que la probabilité d'une telle séquence d'états est donnée par:

$$P(\mathbf{q}|\lambda) = \pi_{q_1} a_{q_1q_2} a_{q_2q_3}\ldots a_{q_{T-1}q_T} \tag{IV.84}$$

La probabilité jointe de $\mathbf{O}$ et $\mathbf{q}$, c'est à dire la probabilité que O et q apparaissent simultanément est simplement le produit de ces deux termes, soit:

$$P(\mathbf{O},\mathbf{q}|\lambda) = P(\mathbf{O}|\mathbf{q},\lambda)\cdot P(\mathbf{q}|\lambda) \tag{IV.85}$$

Ainsi, la probabilité de la séquence $\mathbf{O}$ est obtenue en sommant cette probabilité jointe sur l'ensemble des séquences d'états possibles $\mathbf{q}$, soit:

$$P(\mathbf{O}|\lambda) = \sum_{all\ \mathbf{q}} P(\mathbf{O}|\mathbf{q},\lambda) \cdot P(\mathbf{q}|\lambda) \tag{IV.86}$$

$$= \sum_{q_1,q_2,\ldots q_T} \pi_{q_1} b_{q_1}(\mathbf{o}_1) a_{q_1 q_2} b_{q_2}(\mathbf{o}_2) a_{q_2 q_3} \ldots b_{q_T}(\mathbf{o}_T) a_{q_{T-1} q_T} \tag{IV.87}$$

L'interprétation de ce calcul est le suivant: Au temps $t = 1$, nous sommes dans l'état q_1 avec la probabilité π_{q_1} et générons un symbole $\mathbf{o}_1$ dans cet état avec la probabilité $b_{q_1}(\mathbf{o}_1)$. Le temps passe de t à $t + 1$ (maintenant $t = 2$) et nous faisons une transition à l'état q_2 à partir de l'état q_1 avec la probabilité $a_{q_1 q_2}$, et générons le symbole $\mathbf{o}_2$ avec la probabilité $b_{q_2}(\mathbf{o}_2)$ et ainsi de suite jusqu'à l'observation $\mathbf{o}_T$.

Il est assez simple de montrer que ce calcul s'avère rapidement impossible en raison de sa complexité. En effet, le nombre de calculs nécessaires est égal à $(2T - 1)N^T$ multiplications, et $N^T - 1$ additions.

Il existe cependant une approche connue sous le nom de récurrence avant (*Forward Procedure*) qui permet de résoudre ce problème avec une complexité fortement réduite.

Récurrence avant (*Forward procedure*) Sans entrer dans les détails, l'idée de base de la récurrence avant est de considérer la variable $\alpha_t(i)$ définie comme:

$$\alpha_t(i) = P(\mathbf{o_1},\mathbf{o_2},\mathbf{o_3} \ldots \mathbf{o_t}, q_t = i|\lambda) \tag{IV.88}$$

qui est la probabilité de la séquence d'observations partielle ($\mathbf{o_1},\mathbf{o_2},\mathbf{o_3} \ldots \mathbf{o_t}$ et de l'état i à l'instant t connaissant le modèle λ. La résolution s'effectue à l'aide de la récurrence avant sous la forme:

– Initialisation

$$\alpha_1(i) = \pi_i b_i(\mathbf{o}_1), \qquad 1 \leq i \leq N \tag{IV.89}$$

– Récursion:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_i(\mathbf{o}_{t+1}), \qquad 1 \leq t \leq T-1, \quad 1 \leq j \leq N \tag{IV.90}$$

– Arrêt:

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{IV.91}$$

Notons que cette approche est largement moins complexe puisqu'elle ne demande que $N(N+1)T-1)+N$ multiplications et $N(N-1)(T-1)$ additions. Pour $N = 5$ et $T = 100$, il est assez immédiat de calculer que cette approche est moins complexe (env. 69 ordres de grandeur de différence !!).

De façon similaire, on peut calculer une récurrence arrière en considérant la variable $\beta_t(i)$ telle que:

$$\beta_t(i) = P(\mathbf{o_{t+1}},\mathbf{o_{t+2}},\mathbf{o_{t+3}} \ldots \mathbf{o_T}|q_t = i,\lambda) \tag{IV.92}$$

qui est la probabilité de la séquence d'observations partielle de $t + 1$ à T connaissant le modèle λ et l'état i à l'instant t. La résolution s'effectue à l'aide de la récurrence arrière sous la forme:

– Initialisation

$$\beta_T(i) = 1 \qquad 1 \leq i \leq N \tag{IV.93}$$

– Récursion:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j), \qquad t = T-1, T-2, \ldots, 1, \qquad 1 \leq i \leq N \tag{IV.94}$$

La récursion arrière, associée à la récursion avant permet de proposer une solution aux problèmes 2 et 3 pré-cités.

Problème 2: Retrouver la séquence d'états optimale Contrairement au problème 1 pour lequel on peut obtenir une solution exacte, il s'agit ici plutôt de trouver une séquence optimale connaissant la séquence d'observations.

L'approche couramment retenue est de trouver l'unique meilleure séquence d'états (ou encore chemin) qui maximise la probabilité $P(\mathbf{q}|\mathbf{O},\lambda)$ (ce qui est équivalent à maximiser $P(\mathbf{q},\mathbf{O}|\lambda)$). La méthode permettant de réaliser cela, qui est basée sur la programmation dynamique, est l'algorithme de Viterbi.

Pour cela, on introduit la notion de meilleur chemin partiel jusqu'au temps t et finissant à l'état i. On note $\delta_t(i)$ ce chemin:

$$\delta_t(i) = \max_{q_1, q_2, \cdots q_{t-1}} P[q_1 q_2 \ldots q_{t-1}, q_t = i, \mathbf{o}_1 \mathbf{o}_2, \ldots \mathbf{o}_t | \lambda] \tag{IV.95}$$

Par récurrence, on peut alors déterminer $\delta_{t+1}(i)$ avec :

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \cdot b_j(\mathbf{o}_{t+1}) \tag{IV.96}$$

En pratique, il sera aussi nécessaire de garder la séquence d'états pour chaque t et chaque j. Cela sera réalisé à l'aide du tableau $\psi_t(j)$. On peut ainsi résumer la procédure complète sous la forme:

– Initialisation

$$\delta_1(i) = \pi_i b_i(\mathbf{o}_1), \; 1 \leq i \leq N \tag{IV.97}$$
$$\psi_1(i) = 0 \tag{IV.98}$$

– Récursion:

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \cdot b_j(\mathbf{o}_t), \qquad 2 \leq t \leq T-1; \quad 1 \leq j \leq N \tag{IV.99}$$
$$\psi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}], \qquad 2 \leq t \leq T-1; \quad 1 \leq j \leq N \tag{IV.100}$$

– Arrêt:

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)] \tag{IV.101}$$
$$q_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)] \tag{IV.102}$$

– Rétropropagation (chemin optimal):

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \ldots, 1 \tag{IV.103}$$

Notons qu'en pratique, on utilise souvent un algorithme de Viterbi modifié qui consiste à utiliser le logarithme des paramètres du modèle et ainsi d'éviter d'utiliser des multiplications pour son implémentation (voir [80]) pour plus d'information.

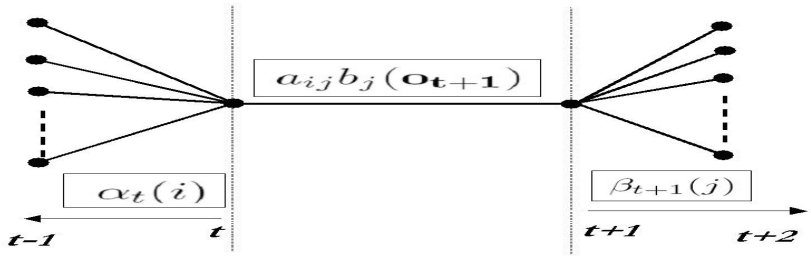


FIG. IV.28 – *sequence d'opérations nécessaire pour le calcul de l'événement joint que le système se trouve dans l'état i à l'instant t et dans l'état j à l'instant $t+1$(d'après [80]).*

Problème 3: Ré-estimer les paramètres du modèle. Le problème de la ré-estimation des paramètres du modèle est plus complexe et peut-être résolu de plusieurs façons différentes. Nous décrivons ci-dessous l'une des procédures itératives, l'algorithme de Baum-Welch (aussi connu sous le nom d'algorithme EM (*Expectation-Maximization*).

L'objectif de la ré-estimation des paramètres du modèle est de choisir un modèle $\lambda = (A,B,\pi)$ tel que sa vraisemblance, $P(\mathbf{O}|\lambda)$ soit maximisée localement.

Soit, $\xi_t(i,j)$, la probabilité d'être dans l'état i à l'instant t et dans l'état j à l'instant $t+1$ connaissant le modèle λ et la séquence d'observation $\mathbf{O}$. La probabilité $\xi_t(i,j)$ s'écrit alors:

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|\mathbf{O},\lambda) \quad \text{(IV.104)}$$

$$= \frac{P(q_t = i, q_{t+1} = j, \mathbf{O}|\lambda)}{P(\mathbf{O}|\lambda)} \quad \text{(IV.105)}$$

Nous pouvons alors écrire cette probabilité comme étant l'ensemble des chemins vérifiant les conditions requises par l'équation IV.105. L'ensemble de ces chemins peut est écrit à l'aide des variables forward et backward (respectivement α_t et β_t définis aux équations IV.88 et IV.92) (voir figure IV.28).

En effet, la sequence d'opérations nécessaire pour le calcul de l'événement joint que le système se trouve dans l'état i à l'instant t et dans l'état j à l'instant $t+1$ correspond à:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(\mathbf{o_{t+1}})\beta_{t+1}(j)}{P(\mathbf{O}|\lambda)}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(\mathbf{o_{t+1}})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(\mathbf{o_{t+1}})\beta_{t+1}(j)} \quad \text{(IV.106)}$$

Soit $\gamma_t(i)$ la probabilité d'être dans l'état i à l'instant t connaissant la séquence d'observation $\mathbf{O}$ et le modèle, alors $\gamma_t(i)$ s'exprime en fonction $\xi_t(i,j)$ en sommant sur l'ensemble des états suivants j:

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \quad \text{(IV.107)}$$

Par ailleurs, si l'on somme $\gamma_t(i)$ sur le temps t, on obtient une quantité qui peut être interprétée comme une estimation du nombre de fois que l'état i est visité. De façon équivalente si l'on ne somme que sur les T-1 premiers indices, cette quantité peut être interprétée comme

une estimation du nombre de transitions à partir de l'état i. De même, si l'on somme la variable $\xi_t(i,j)$ sur le temps (de $t=1$ à $t=T-1$), on peut interpréter cette nouvelle quantité comme l'estimation du nombre de transitions de l'état i vers l'état j.

Ainsi, nous avons:

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{estimation du nombre de transitions à partir de l'état i dans } \mathbf{O} \tag{IV.108}$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{estimation du nombre de transitions de l'état } i \text{ vers l'état } j \text{ dans } \mathbf{O} \tag{IV.109}$$

On peut ainsi en déduire, à partir des formules précédentes une méthode pour ré-estimer les paramètres d'un modèle de Markov caché. Nous avons ainsi les formules de ré-estimation suivantes:

$$\begin{aligned}
\overline{\pi}_j &= \text{estimation du nombre de fois dans l'état } i \text{ à l'instant } t=1 \\
&= \gamma_1(i) && \text{(IV.110)} \\
\overline{a}_{ij} &= \frac{\text{estimation du nombre de transitions de l'état } i \text{ à l'état } j}{\text{estimation du nombre de transitions à partir de l'état } i} && \text{(IV.111)} \\
&= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} && \text{(IV.112)} \\
\overline{b}_j(k) &= \frac{\text{estimation du nombre de fois dans l'état } j \text{ en y observant le symbole } \mathbf{v_k}}{\text{estimation du nombre de fois dans l'état } j} \\
&= \frac{\sum_{t=1,\mathbf{o_t}=\mathbf{v_k}}^{T-1} \gamma_t(i)}{\sum_{t=1}^{T-1} \gamma_t(i)} && \text{(IV.113)}
\end{aligned}$$

Les équations ci-dessous permettent à partir du modèle courant $\lambda=(A,B,\pi)$ de ré-estimer les paramètres et d'obtenir un modèle ré-estimé $\overline{\lambda}=(\overline{A},\overline{B},\overline{\pi})$. Il a été montré par Baum et ses collègues que soit le modèle initial λ définit un minimum local de la fonction de vraisemblance et auquel cas, $\overline{\lambda}=\lambda$, soit le modèle $\overline{\lambda}$ est plus probable que le modèle λ c'est à dire que $P(\mathbf{O}|\overline{\lambda}) > P(\mathbf{O}|\lambda)$. Nous avons ainsi trouvé un nouveau modèle $\overline{\lambda}$ pour lequel la séquence d'observation est plus probable qu'avec l'ancien modèle λ.

Le lecteur intéressé pourra consulter [80] ou [15] pour plus de détails et pour la description de l'approche (équivalente) utilisant des techniques d'optimisation classique pour la maximisation par rapport à λ de la fonction auxiliaire de Baum $Q(\lambda',\lambda)=\sum_{\mathbf{q}} P(\mathbf{O},\mathbf{q}|\lambda')\log P(\mathbf{O},\mathbf{q}|\lambda)$.

IV.6.3 Densités d'observation continues

Dans tout ce qui précède, nous avons supposé que les observations étaient caractérisés par des symboles discrets, c'est à dire prenant des valeurs dans un ensemble fini de valeurs possibles. Dans un tel modèle, on pouvait donc associer des densités de probabilité d'observation discrète pour chaque état de notre modèle. Cependant, pour les applications de traitement de la parole, les observations (vecteurs cepstraux) ne sont pas en général discrètes. Il est ainsi avantageux de pouvoir utiliser dans notre modèles des densités de probabilité d'observation continues. On ne

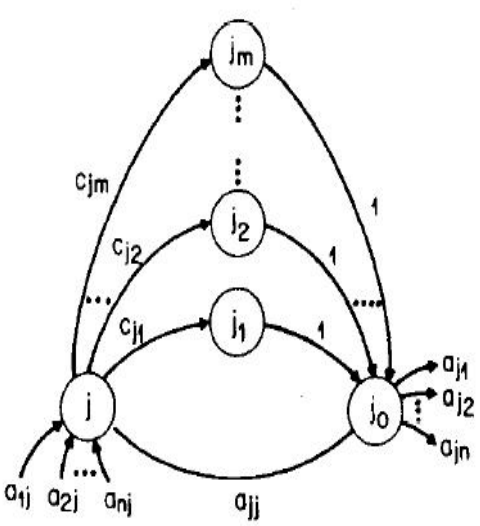


FIG. IV.29 – *Equivalence entre un modèle à un état avec une densité sous forme d'un mélange de Gausiennes et d'un modèle multi-états en parallèle avec des densité sous forme d'une Gausienne (d'après [80]).*

peut malheureusement pas considérer n'importe quelle densité de probabilité et il faut en général se doter d'un modèle pour lequel on peut obtenir des formules de réestimation . Le modèle le plus couramment utilisé pour représenter la densité de probabilité d'observation et pour lequel il existe des formules analytiques de réestimation est le mélange de Gaussiennes:

$$b_j(\mathbf{o}) = \sum_{k=1}^{M} c_{jk} \aleph(\mathbf{o}, \mu_{jk} \Gamma_{jk}), \quad 1 \leq j \leq N \tag{IV.114}$$

où $\mathbf{o}$ est le vecteur d'observation, c_{jk} le coefficient correspondant à la k^{ieme} Gaussienne de la densité de probabilité de l'état j et où $\aleph$ est une Gaussienne de moyenne μ_{jk} et matrice de covariance Γ_{jk}. Les coefficients c_{jk} doivent vérifier la contrainte:

$$\sum_{k=1}^{M} c_{jk} = 1, \quad 1 \leq i \leq N \tag{IV.115}$$

$$c_{jk} \geq 0, \quad 1 \leq j \leq N, \quad 1 \leq k \leq M \tag{IV.116}$$

ce qui permet d'avoir des densités de probabilité d'observation normalisées pour chaque état:

$$\int_{-\infty}^{+\infty} b_j(\mathbf{o}) d\mathbf{o} = 1 \quad 1 \leq j \leq N \tag{IV.117}$$

Il est intéressant ici de constater qu'un état d'une chaîne de Markov cachée avec une densité d'observation sous la forme d'un mélange (ou somme) de Gausiennes est équivalent à un modèle multi-états en parallèle contenant chacun une des Gaussiennes de la mixture (voir figure IV.29).

Il peut être montré que les formules de réestimation dans le cadre d'un modèle de somme de Gaussiennes sont de la forme:

$$\overline{c_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{IV.118}$$

$$\overline{\mu_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k).\mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{IV.119}$$

$$\overline{\mathbf{\Gamma}_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k).(\mathbf{o}_t - \mu_{jk})(\mathbf{o}_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{IV.120}$$

où $\overline{c_{jk}}$, $\overline{\mu_{jk}}$ et $\overline{\mathbf{\Gamma}_{jk}}$ sont les paramètres du mélange de Gaussiennes et où $\gamma_t(j,k)$ est la probabilité d'être dans l'état j au temps t avec la k^{ieme} Gaussienne du mélange représentant $\mathbf{o}_t$ et est donnée par:

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}\right]\left[\frac{c_{jk}\aleph(\mathbf{o_t},\mu_{jk}\Gamma_{jk})}{\sum_{m=1}^{M} c_{jm}\aleph(\mathbf{o_t},\mu_{jm}\Gamma_{jm})}\right] \tag{IV.121}$$

Les formules de réestimation pour les a_{ij} sont identiques aux formules pour des observations discrêtes (voir equations IV.111). L'interprétation des formules de réestimation est ici assez directe. Par exemple, l'équation IV.118 est le rapport entre le nombre estimé de fois où l'on est dans l'état j en utilisant la k^{ieme} Gaussienne et le nombre estimé de fois où l'on est dans l'état j.

IV.7 Vers la reconnaissance robuste

La miniaturisation d'un grand nombre d'appareils électroniques (ou terminaux) aux fonctions de plus en plus avancées a rendu leur utilisation difficile. Les technologies vocales, et notamment la reconnaissance vocale, permet d'avoir une interface plus simple et plus intuitive pour ces terminaux. La technologie est par ailleurs maintenant tout à fait performante dans des environnements contrôlés. Néanmoins, comme l'utilisation de ces terminaux est faite de plus en plus dans des environnements dits difficiles (bruit, distorsion due au canal, réverbération,...) les performances de reconnaissance vocale s'en retrouvent considérablement affectées. L'objet de ce chapitre est ainsi de présenter quelques approches qui permettent de mieux contrôler ces dégradations. Pour plus de détails, on pourra se référer à [87] à partir duquel cette section a été rédigée.

IV.7.1 Effet du bruit sur les systèmes de reconnaissance

L'effet du bruit sur le signal revient à transformer le signal propre s (i.e. le signal de parole produit en l'absence de bruit) en un signal bruité s_b. En notant $T_n()$ la transformation due à cette corruption, nous avons:

$$s_b = T_n(s) \tag{IV.122}$$

Notons que la transformation peut représenter un grand nombre de situations allant de la simple addition d'un bruit jusqu'aux transformations linéaires ou non-linéaires plus complexes telles que la convolution, la compression, la saturation, etc...

Avec ces notations, on peut écrire que les paramètres (i.e par exemple les MFCC) calculés à partir du signal bruité s_b sont alors donnés par $F(s_b) = F(T_n(s))$. Ainsi, idéalement, un système de reconnaissance opérant sur des signaux bruités serait aussi entraîné sur de la parole bruitée car de fait les modèles appris sur de la parole propre diffèrent des modèles appris sur de la parole bruitée puisqu'ils n'intègrent pas la transformation T_n mentionnée ci-dessus.

Cependant, en pratique les conditions de bruit (et donc la transformation T_n affectant le signal de parole) peuvent être de nature très variée et peuvent également varier au cours du temps. Ainsi, dans de nombreuses situations, il n'est pas aisé voire impossible de prédire ou de déterminer de façon exacte la transformation T_n entrant en jeu.

En l'absence d'informations *a priori* fiables, il est ainsi courant d'entraîner le système de reconnaissance sur des signaux sensés représenter l'ensemble des situations d'utilisation. Cette solution n'est malheureusement pas souvent satisfaisante car même lorsque la transformation est relativement stable pour les données d'apprentissage utilisées les performances de reconnaissance chutent. En effet, lorsque le niveau de bruit augmente, les signaux ressemblent de plus en plus à du bruit et les distinctions entre les différents sons de la parole en sont ainsi diminuées. La conséquence est qu'un système de reconnaissance optimal robuste au bruit sera toujours moins performant qu'un système optimal de reconnaissance de parole propre.

Il est important aussi de noter qu'il y a différents types de bruit et qu'ils affectent les performances de reconnaissance de façon différente. On sépare souvent les bruits stationnaires (bruit blanc) des bruit non-stationnaires tels que la musique. Contrairement à ce que l'on pourrait penser, les bruits stationnaires sont souvent beaucoup plus perturbateurs que les bruits non-stationnaires. La raison principale est qu'un bruit non-stationnaire à Rapport Signal à Bruit (RSB) donné possède de larges portions du signal pour lequel le RSB instantané est bien supérieur due à la nature variable du bruit. Par contre, il est en général beaucoup plus compliqué de réduire l'impact de tels bruits comparé à la réduction des bruits stationnaires

IV.7.2 Compenser l'effet du bruit

Comme on vient de le voir, la présence de bruit a pour conséquence directe une réduction des performances de reconnaissance. Une réduction de cette dégradation pourra être obtenue à condition de parvenir à réduire la différence entre la distribution actuelle des données de test (bruitées) et celle des données d'apprentissage (en général donc non-bruitées).

Pour reprendre les notations de la section précédente, il s'agit de minimiser la dégradation en performances due au bruit, c'est à dire minimiser la différence entre $P(F(s_b)|\lambda)$ et $P(F(s)|\lambda)$ qui représentent respectivement les probabilités des observations pour le signal bruité et le signal non-bruité connaissant le modèle λ (qui peut être ici le modèle correspondant à un phonème ou mot). Il faut bien sur minimiser cette différence pour tous les modèles. Cela peut être effectué de plusieurs façons qui sont brièvement résumées ci-dessous:

Débruitage du signal Dans cette approche, il s'agit d'appliquer une transformation $C()$ directement sur le signal dégradé ou bruité s_b, de telle sorte que $z = C(s_b) \approx s$. La reconnaissance est alors effectuée sur les paramètres $F(z)$ extraits du signal transformé ou débruité z. On comprend que plus la transformation est efficace et que le signal z est proche du signal propre cible s, plus les probabilités $P(F(z)|\lambda)$ et $P(F(s)|\lambda)$ seront proches et donc plus les performances de reconnaissance seront équivalentes à celles obtenues sur un signal propre (à fort RSB). Pour ce type de méthodes, le lecteur est renvoyé au chapitre "Débruitage" du cours MSA de la brique PAMU. On trouvera aussi une approche développée plus spécifiquement dans un but de reconnaissance vocale dans [60] et la solution retenue par le standard ETSI-AURORA [91].

Compensation de caractéristiques Cela est réalisé en transformant les caractéristiques calculées sur la parole bruitée, $F(s_b)$, avec une transformation $C()$, telle que $f_z = C(F(s_b)) = F(s)$. La reconnaissance est alors effectuée sur les paramètres transformés f_z en lieu et place des paramètres bruts extraits du signal bruité $F(s_b)$. On comprend que plus la transformation est efficace, plus les probabilités $P(f_z)|\lambda)$ et $P(F(s)|\lambda)$ seront proches et donc plus les

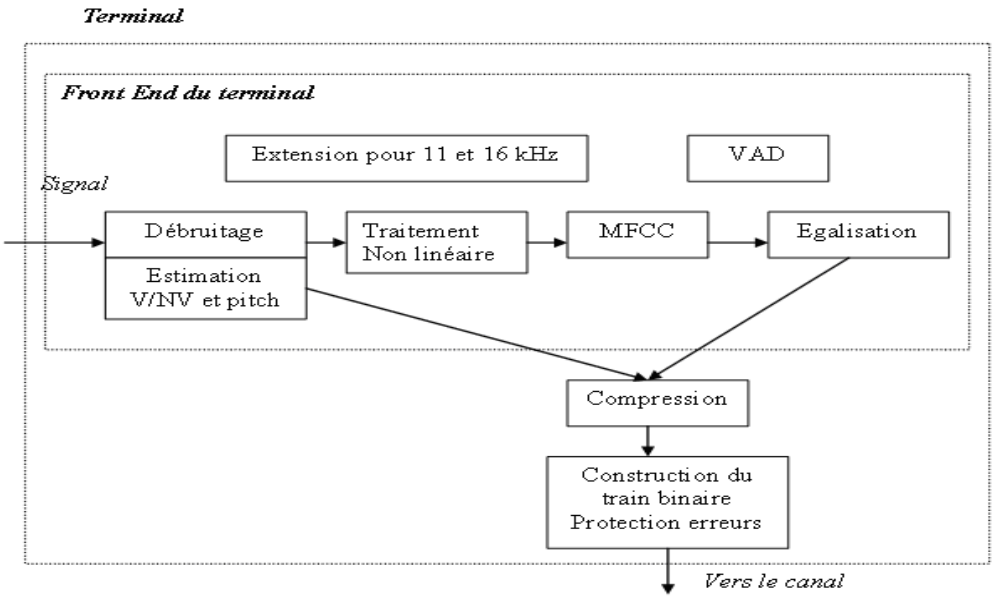


FIG. IV.30 – *Schéma bloc du front end étendu du standard Aurora (coté terminal)*

performances de reconnaissance seront équivalentes à celles obtenues sur un signal propre.

Compensation de modèles Cela est réalisé en transformant les distributions apprises à l'apprentissage de façon à mieux représenter les données à reconnaître. Cette approche est évidemment dépendante du modèle utilisé contrairement aux deux approches précédentes. Cette approche peut être formalisée en disant que l'on cherche une transformation $C()$ appliquée aux distributions de probabilité telle que $C(P(F(s))|\lambda) \approx P(F(s_b)|\lambda)$. La reconnaissance est ici faite en utilisant les nouvelles distributions $C(P(F(s))|\lambda)$.

Techniques d'adaptation Dans ces approches, l'objectif est de faire correspondre les distributions utilisées par le système de reconnaissance à celles des données de test. En apprentissage adapté (*matched training*), cela est effectué en entraînant le système de reconnaissance sur les données bruités en essayant d'utiliser différentes conditions de bruit qui correspondent à celles qui seront rencontrées en condition de test. Dans d'autres approches apparentées à la théorie des données manquantes (*missing feature theory*), l'objectif est de n'utiliser que les segments de données dont les distributions de paramètres correspondent aux distributions modélisées par le système de reconnaissance à l'apprentissage et de traiter les autres segments (*mismatched components*) comme des données manquantes.

IV.7.3 Le standard ETSI-Aurora

Le standard ETSI-Aurora a été spécifié pour une architecture distribuée de la reconnaissance vocale (DSR, pour *Distributed Speech Recognition*). Une telle architecture permet de mieux lutter contre les erreurs de transmissions (particulièrement importantes pour les communications mobiles) et les dégradations dues au codage de source (et en particulier à la limitation de la bande à 4kHz). Ainsi, dans une telle structure le terminal effectue l'extraction de paramètres (encore appelé *front end*), qui sont ensuite transmis sur le canal au système de reconnaissance central (encore appelé *back end*). Ainsi, il n'existe plus de dégradations apportées par une opération de codage/décodage de la parole et le canal peut

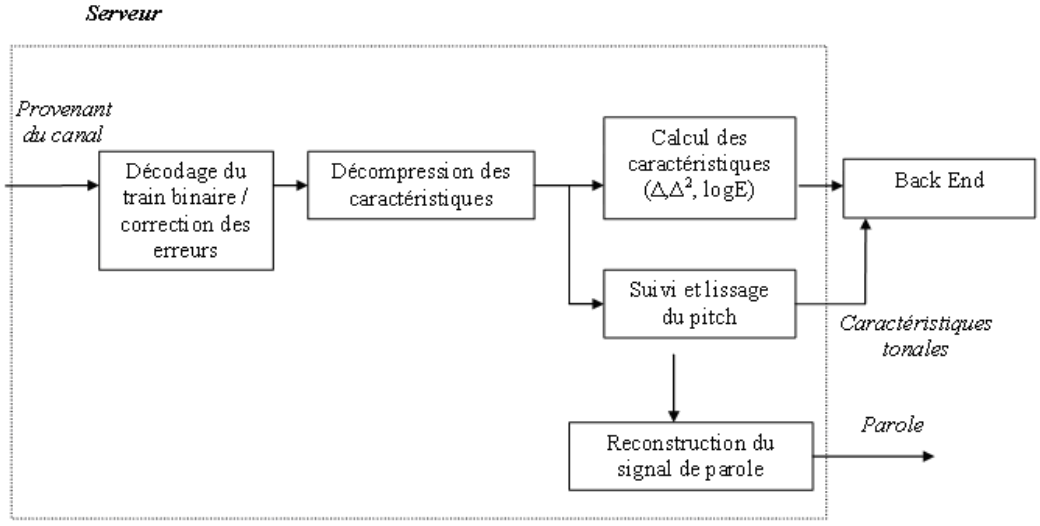


FIG. IV.31 – *Schéma bloc du front end étendu du standard Aurora (coté serveur)*

être considéré comme invariant.
Il existe en fait deux versions du standard ETSI Aurora. La première version publiée en février 2000 était basée sur l'utilisation des coefficients MFCC qui sont très couramment utilisées en reconnaissance vocale. La seconde version qui est une version étendue de la première est présentée dans ses grandes lignes ici (dénommé *Advanced DSR Front End*) et permet d'obtenir des performances significativement supérieures notamment en présence de bruit. On donne ci-dessous quelques précisions sur les différents éléments constituant le standard Aurora mais pour plus de détails, on pourra se référer directement au document spécifiant le standard [91]. Le schéma général du standard est présenté sur les figures IV.30 et IV.31. De façon générale, le standard Aurora calcule une paramétrisation du signal de type coefficients MFCC après débruitage du signal par un filtrage de Wiener sur échelle Mel et avant compression pour transmission sur le canal. Il consiste aussi à estimer des paramètres supplémentaires et notamment à tracker la fréquence fondamentale pour pouvoir reconstruire, au niveau du serveur, un signal de parole (pour par exemple vérifier la qualité de la parole pour les phases d'apprentissage ou d'adaptation du système de reconnaissance). On remarquera aussi que le standard inclut une extension pour la prise en compte des signaux échantillonnés à 11 ou 16 kHZ. On donne ci-dessous quelques précisions sur quelques éléments constituant le standard Aurora mais pour plus de détails, on pourra se référer directement au document décrivant le standard où chaque bloc est très précisément spécifié [91].
Le schéma général du standard est présenté sur les figures IV.30 et IV.31.

Débruitage

Le signal est tout d'abord débruité à l'aide d'une approche itérative basée sur le filtre de Wiener. Le principe de base consiste à, dans un premier temps, débruiter le signal avant d'opérer une nouvelle étape de débruitage dépendant du rapport Signal à Bruit du signal.

Le filtre de Wiener est ici calculé après transformation sur une échelle Mel. La présentation détaillée de cette approche dépasse le cadre de ce cours.

Traitement non-linéaire

Ce module de traitement non linéaire (appelé *Waveform processing* dans le standard) consiste à ne retenir qu'une partie du signal débruité s_d pour les étapes suivantes de calcul. Plus précisément, l'objectif est ici de pondérer le signal débruité s_d par une fenêtre w définie autour des maxima de l'enveloppe du signal considéré.
La première étape de module consiste à calculer une enveloppe lissée $e(n)$. Pratiquement, cette enveloppe lissée est calculée en effectuant une moyenne des enveloppes instantanées obtenues à l'aide de l'opérateur introduit par Teager $e(n) = \frac{1}{9}\sum_{i=-4}^{i=4} e_{inst}(n+i)$ avec $e_{inst}(n) = |s^2(n) - s(n-1) * s(n+1)|$. A partir de cette enveloppe lissée, une étape de détection de maxima est effectuée itérativement. Le maximum global de l'enveloppe sur chaque fenêtre de 240 échantillon est sélectionné puis les maxima voisins sont sélectionnés en prenant soin de ne conserver que les maxima qui sont distants de 25 à 80 échantillons de ses voisins immédiats. A partir du nombre de maxima N_{max} de l'enveloppe lissé $e(n)$ et leurs positions $P_{max}(n_{max})$ pour $0 \leq n_{max} < N_{max}$, on définit une fonction de pondération w sous la forme:

$$w = \begin{cases} 1 & \text{pour n} \in [n_{inf}; n_{inf} + 0.8 \times (P_{max}(n_{max}+1) - P_{max}(n_{max}))] \\ 0 & \text{sinon} \\ 0.5 & \text{aux transitions entre 0 et 1} \end{cases} \quad \text{(IV.123)}$$

où $n_{inf} = P_{max}(n_{max}) - 4$.
On applique ensuite cette fenêtre de pondération au signal débruité s_d :

$$s_w(n) = 0.8 * s_d(n) + 0.4 * s_d(n) * w(n) \quad \text{(IV.124)}$$

IV.7.4 Compensation de caractéristiques (normalisation cepstrale, RASTA)

On trouve un grand nombre d'approches permettant de réaliser une compensation de caractéristiques. Certaines méthodes utilisent la connaissance explicite du bruit environnant qui est enregistré simultanément au signal bruité (bruit + signal bruité) ou au signal propre, par exemple enregistré à l'aide d'un micro-casque (bruit + signal propre). On parle ici de données "stéréo". La reconnaissance de parole en voiture lorsque l'autoradio est allumé peut par exemple exploiter la connaissance du signal produit par la radio pour compenser les caractéristiques extraites sur le signal bruité.
Une autre utilisation possible consiste à apprendre les relations statistiques entre les caractéristiques des données bruités et celles des données propres qui sont ensuite utilisées pour compenser l'effet du bruit sur les données de test (la méthode RATZ [87] entre dans ce cadre). D'autres types d'algorithmes, dénommés paramétriques, n'auront accès qu'aux données bruitées et à la connaissance des modèles calculées sur les données propres. Pour ces approches, l'objectif sera de caractériser l'effet du bruit sur les caractéristiques à l'aide d'un modèle paramétrique. Ce modèle est ensuite utilisé avec les données de test bruitées et les distributions connues des caractéristiques des signaux propres pour estimer les paramètres statistiques qui représentent le bruit et qui seront utilisés pour compenser les caractéristiques des signaux bruités

Normalisation cepstrale

Cependant, l'une des approches les plus populaires en raison de sa simplicité mais aussi de ses performances est un simple filtrage passe-haut des coefficients cepstraux. C'est ce que fait la normalisation par la moyenne cepstrale (*Cepstral mean normalisation*) où la moyenne des coefficients cepstraux est soustraite de tous les vecteurs cepstraux. Ainsi si c_t représentent les N coefficients cepstraux au temps t, les nouveaux coefficients compensés notés $\tilde{c}_t$ seront donnés par $\tilde{c}_t = c_t - \frac{1}{T}\sum_{t=1}^{T} c_t$ où T est la durée sur laquelle est calculée la moyenne.
Il existe d'autres approches de type filtrage, et notamment les méthodes connues sous le nom de RASTA (pour *RelAtive SpectrAl*) et J-RASTA [45],[46] qui sont souvent reconnues comme plus robustes qu'une simple normalisation cepstrale.

Méthodes RASTA et J-RASTA

La principale motivation des techniques RASTA et J-RASTA est de filtrer les variations spectrales qui sont soit trop rapides pour correspondre à un mouvement d'un des articulateurs du conduit vocal soit trop lentes et qui seraient donc plutôt dues à un effet du canal ou d'un bruit de fond. L'analyse RASTA consiste en règle générale à la succession de 3 étapes clés à partir d'une représentation spectrale en bandes critiques ou MEL:

- Transformation non linéaire de type compression (dans la version log-RASTA une simple fonction logarithmique est utilisée).
- Filtrage des trajectoires temporelles de chaque composante spectrale transformée
- Transformation non-linéaire inverse de type expansion (dans la version log-RASTA une simple fonction exponentielle est utilisée).

Le filtre proposé dans la méthode originale ([46]) a pour fonction de transfert:

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{IV.125}$$

Une implémentation causale de ce filtre sera donnée par l'équation aux différences suivante:

$$y(n) = 0.98y(n-1) + 0.2x(n) + 0.1x(n-1) - 0.1x(n-3) - 0.2x(n-4) \tag{IV.126}$$

Si l'on considère que ce filtre est appliqué sur des signaux sous-échantillonnés à 100 Hz (ce qui correspond à 100 fenêtres d'analyse par seconde soit un décalage entre deux fenêtres d'analyse successives de 10 ms), cela nous donne une fréquence de coupure basse à 0.26 Hz, une pente spectrale de 6 dB/oct à partir de 12.8 Hz avec des zéros à 28.9 Hz et 50 Hz (voir figure IV.32). Ici, le rôle du zéro à 28.9 Hz n'apparaît pas clairement mais pourrait être lié à la longueur des fenêtres d'analyse utilisées.
La méthode J-RASTA consiste à remplacer l'étape de compression de l'approche log-RASTA par la fonction :

$$y = \ln(1 + Jx) \tag{IV.127}$$

où J est une constante positive qu'il convient d'optimiser. Il peut être montré que la valeur optimale de J dépend du rapport Signal à Bruit du signal de parole considéré. On constate que pour $J \ll 1$ la transformation est quasi-linéaire alors que pour que $J \gg 1$, la transformation est quasi-logarithmique et on se rapproche ici de la méthode RASTA

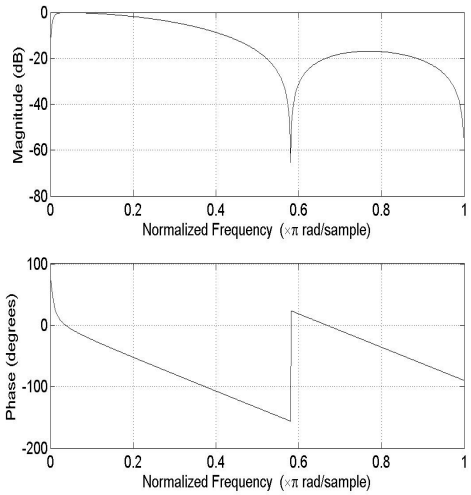


FIG. IV.32 – *Gain et phase du filtre initialement proposé par Hermansky pour la méthode RASTA*

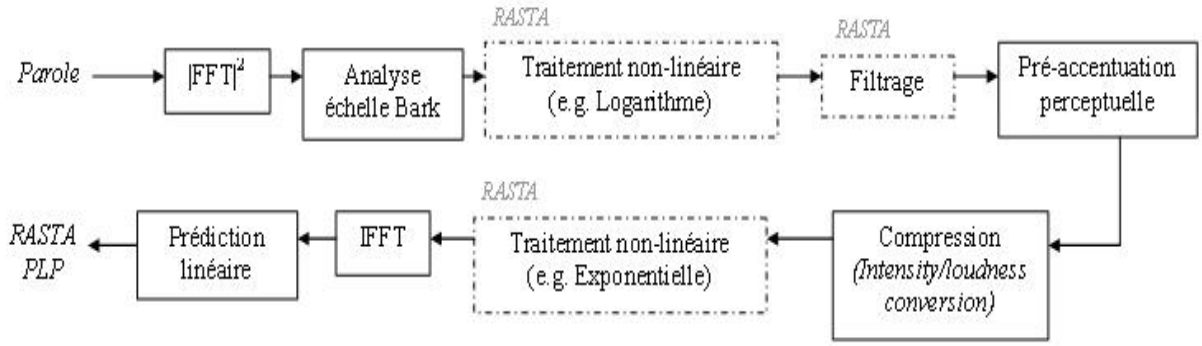


FIG. IV.33 – *Principales étapes pour le calcul des RASTA-PLP (les étapes concernant la méthode RASTA sont indiquées en pointillé)*

simple. Initialement proposée en association avec la paramétrisation PLP, les méthodes RASTA peuvent bien sur être implémentées avec une paramétrisation MFCC.
La figure IV.33 donne le schéma général des blocs de traitement qui sont ajoutés au calcul des paramètres PLP pour donner lieu aux paramètres RASTA-PLP.

Compensation de modèles

De même, de nombreuses méthodes existent incluant notamment la combinaison de modèles parallèles ([41]), l'adaptation MLLR ou encore l'adaptation par Maximum A posteriori [42]. Ces méthodes ne seront pas détaillées dans ce document. Le lecteur intéréssé pourra cependant consulter [86].

IV.8 Bases de données pour la reconnaissance vocale

IV.8.1 Introduction

Les technologies de reconnaissance de parole sont maintenant matures et permettent de développer des interfaces vocales pour de très nombreuses applications. Les moteurs de reconnaissance sont dans la plupart des cas (et en particulier lorsqu'ils opèrent en mode indépendant du locuteur) construits à partir de modèles de Markov cachés (HMM). Les modèles de Markov cachés doivent être entraînés (ou appris) à l'aide de bases de données qui reflètent au mieux la variabilité des locuteurs et des environnements acoustiques concernés par l'application. Ainsi, ces bases de données incluent un grand nombre de locuteurs. Pour être utiles à la reconnaissance de parole (apprentissage), ces bases de données doivent également être annotées, c'est-à-dire apporter une description précise du contenu de la base.

C'est pour répondre à ces besoins, qu'une structure internationale fut créée pour construire ces bases de données puis pour les diffuser. Au sein des projets de recherche américain en parole financés par DARPA, un certain nombre de bases de données particulièrement utiles pour la recherche virent le jour. Ces bases sont aujourd'hui distribuées par LDC (Linguistic Data Consortium [20]). Plus récemment, le besoin d'avoir une structure en Europe dont le but serait de distribuer des bases de données aussi bien pour la recherche que pour l'industrie s'est fait sentir et a ainsi amené la création d'ELRA (European Linguistic Ressources Association [4]). Dans le but de produire des bases de données directement exploitables par l'industrie, plusieurs projets européens ont été lancés (les projets de la famille SpeechDat [78] et plus récemment le projet Speecon [77].

Sans décrire toutes ces bases de données, il est important de noter que pour obtenir des systèmes de reconnaissance performants dans des conditions d'utilisation données, il est nécessaire d'enregistrer un grand nombre de données dans ces environnements cibles. Ainsi, le projet SpeechDat-Car qui vise les applications automobiles enregistre les données en voiture et le projet Speecon qui s'intéresse à toutes les applications " grand public " (jouets, télévision, assistant personnel, téléphone mobile,..) enregistre les données dans ces situations d'utilisation.

IV.8.2 Un exemple: la base de données SpeechDat-Car

Le but du projet *SpeechDat-Car* est de construire des bases de données qui permettront de développer des algorithmes de reconnaissance de parole plus robustes pour de nombreuses applications automobiles et ce pour une dizaine de langues.

De fait, l'émergence d'accessoires de plus en plus nombreux dans une voiture (autoradio, téléphone, systèmes de navigation,.) procure au conducteur d'une voiture moderne de nouvelles fonctionnalités mais en même temps le place dans une situation difficile puisque l'utilisation de ces accessoires le distrait très clairement de sa tâche principale qui est de conduire. C'est pourquoi la reconnaissance de parole apparaît comme une technologie particulièrement adaptée pour à la fois permettre le déploiement de ces applications (grâce à une utilisation " mains-libres ") et garantir un haut niveau de sécurité. Cependant, pour obtenir des performances de reconnaissance optimales dans un " environnement voiture " (qui est connu pour être très bruité), il est nécessaire d'entraîner le système de reconnaissance en utilisant de grandes bases de données enregistrées en contexte (i.e. directement dans la voiture).

Afin d'obtenir les bases de données multilingues, plusieurs phases sont nécessaires:

- *Les spécifications:* cette phase vise à définir précisément:
  - les spécifications de contenu, c'est à dire le vocabulaire utilisé, les phrases et mots

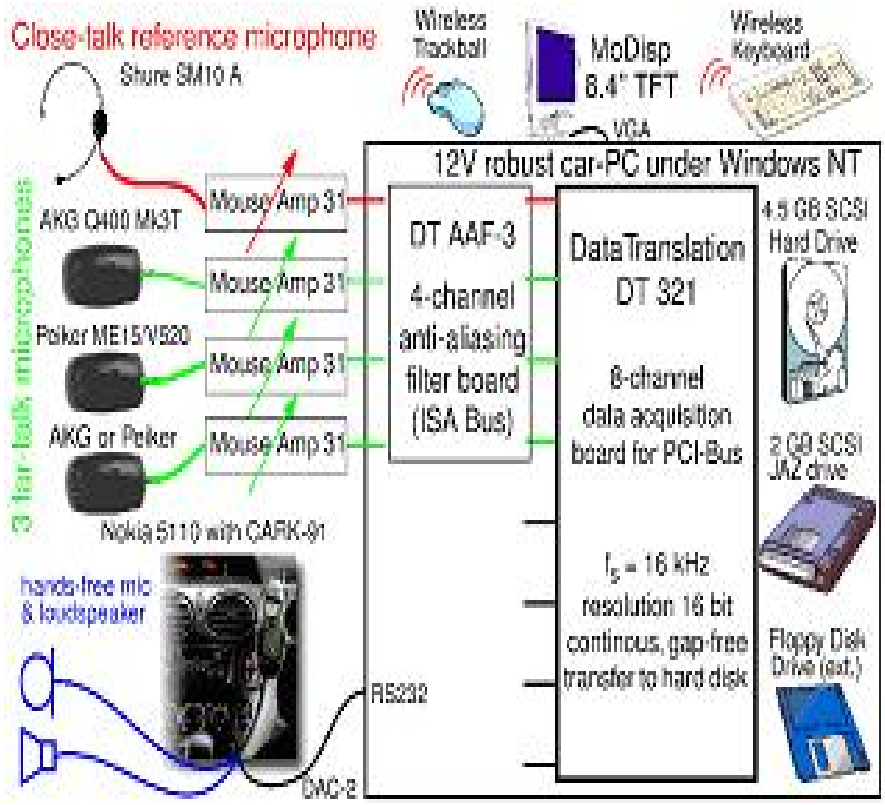


FIG. IV.34 – *Schéma de la plateforme d'enregistrement pour le projet SpeechDat-Car(d'après [39]).*

à prononcer, le style d'élocution (parole lue, parole spontanée, etc...), le nombre de locuteurs, les conditions d'environnement (voiture à l'arrêt, sur autoroute, en ville, etc...)

- les spécifications de matériels (plate-formes d'enregistrements incluant le type de microphones utilisées, etc...). La figure IV.34 donne un schéma de la platteforme utilisée dans le projet SpeechDat-Car
- les spécifications de format d'enregistrement (noms des fichiers, arborescence, description des paramètres d'enregistrement, etc.). Par exemple, pour l'ensemble des projets SpeechDat, il a été décidé d'utiliser le format SAM, et d'enregistrer ainsi ces informations dans un fichier texte séparé du fichier audio correspondant.

Pour plus de détails, on pourra consulter [61],[24] ou [39].

- *L'enregistrement:* Cette phase consiste en l'enregistrement proprement dit des signaux. Cette phase n'est pas nécessairement triviale surtout lorsque les conditions d'enregistrement sont difficiles, comme cela peut être le cas pour une base réalisée en voiture. Pour SpeechDat-Car, chaque locuteur avait un certain nombre de phrases/mots à prononcer sachant que chaque session durait environ 1 heure. Le contenu d'une session d'enregistrement est donné sur la figure IV.36
- *L'annotation:* cette phase vise à annoter les signaux enregistrés c'est à dire à attacher une information textuelle aux signaux enregistrés. Cette annotation peut être très précise (par exemple, aller jusqu'à une annotation phonétique). Pour SpeechDat-Car, une annotation

SAM Label	Description	SRC	speech file name
LHD	label header	CCD	corpus code
ELF	end of label file	REP	recording place
CMT	comment	RED	recording date
DBN	database name	RET	recording time
SES	session number	BEG	labelled sequence begin position
REG	calling regio	END	labelled sequence end position
NET	network	SAM	sampling frequency
PHM	phone model	SNB	number of (8-bit) bytes/sample
SCD	speaker code	SBF	sample byte order
SEX	speaker gender	SSB	number of significant bits/sample
AGE	speaker age	QNT	quantization
ACC	speaker accent	NCH	number of channels
DIR	speech file directory	LBD	label file body
SRC	speech file name	LBR	prompt text
MIP	microphone position	MIT	microphone type

FIG. IV.35 – *Quelques uns des labels SAM utilisés pour la base de données SpeechDat-Car(d'après [39]).*

Items	# per session	Tot. expected number of repetitions
Isolated digit	4	200 per digit
Digit strings		-
10 digits in isolation	1	600 per digit
Telephone number	3	-
Spontaneous tel number	1	-
Credit card number (set of 150)	1	4 per number
PIN code (set of 150)	1	4 per number
Sheet number	1	-
Natural numbers	1	-
Money amounts	1	-
Dates		
Spontaneous	1	-
Absolute	1	50 per month name; 85 per dayname
Relative (set of 10)	1	60 per date
Times		
Spontaneous	1	-
Analog	1	-
Names		
Birth city (spontaneous)	1	-
Most important cities (set of 150)	2	8 per name
Most important companies (set of 150)	2	8 per name
Forename or surname (spontaneous)	1	-
Forename + surname (set of 150)	1	4 per name
Spelling		
Forename/surname	1	-
Word/name	4	-
Artificial word	1	600 per letter
City name	1	-
Phonetically rich words	4	200 per phone
Phonetically rich sentences	9	250 per phone
Application words (set of 201)	67	200 per word
Spont. phrase with application word	2	-
Voice activation keywords (set of 5)	2	240 per word
Additional language-dependent application words (set of 10)	2	120 per word
Spontaneous sentences (for the last 100 speakers)	10	90 per situation
TOTAL	**119**	**(for the first 200 speakers)**
	129	**(for the last 100 speakers)**

FIG. IV.36 – *Contenu d'une session d'enregistrement de la base SpeechDat-Car (d'après [39]).*

graphémique est donnée et elle inclut ainsi le texte qui a été prononcé par le locuteur accompagné de quelques marqueurs généraux (bruit de bouche, hésitations etc.). Cette phase, assez fastidieuse, ne doit pas être négligée car c'est elle qui permet d'utiliser la base de données pour entraîner les systèmes de reconnaissance. (Pour plus de détails concernant l'annotation, on pourra consulter [25])).

- *La documentation:* Cette phase vise à construire une documentation décrivant le contenu de la base de données en fournissant notamment des statistiques de la base (nombre d'occurrences de chaque mots, proportion effective homme/Femme, etc ...).
- *La validation:* Cette phase, pourtant essentielle pour pouvoir évaluer la qualité d'une base de données, est trop souvent négligée. Idéalement, la validation d'une base de données sera réalisée par un centre indépendant et visera entre autres à vérifier la qualité de l'annotation et la qualité d'enregistrement des signaux. Cette validation peut être effectuée en écoutant un nombre d'occurrences qui auraient été tirées au hasard. On peut également utiliser un certain nombre de critères objectifs tels que le rapport signal à bruit, le taux de saturation des signaux, etc .... Il est à noter qu'une phase de pré-validation peut être avantageusement organisée à partir de l'enregistrement d'un nombre limité de locuteurs (par exemple 6 locuteurs) afin de vérifier que les spécifications sont bien suivies et qu'il n'y a pas de problèmes importants concernant la qualité des signaux. Une telle phase de pré-validation est organisée dans le cadre des projets SpeechDat-Car et permet de minimiser le risque d'enregistrer une base qui s'avérerait inutile ou trop mauvaise pour la reconnaissance de parole.

Par ailleurs, on trouvera dans [39] une description détaillée du projet ainsi qu'une analyse concernant les données dites spontanées où il est montré que les situations entièrement spontanées (requête émise par le locuteur) diffèrent des situations semi-spontanées (réponse du locuteur à une question simple de l'opérateur) pour lesquelles un nombre très faible d'hésitations est constaté. On pourra par ailleurs consulter le site WEB du projet SpeechDat-Car [76].

Chapitre V

Synthèse de la parole[1]

V.1 Définition

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

V.2 Architecture d'un système TTS

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

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

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

1. Chapitre reprenant de larges extraits du polycopié de cours de T. Dutoit [27] et du cours de F. Beaugendre [10]

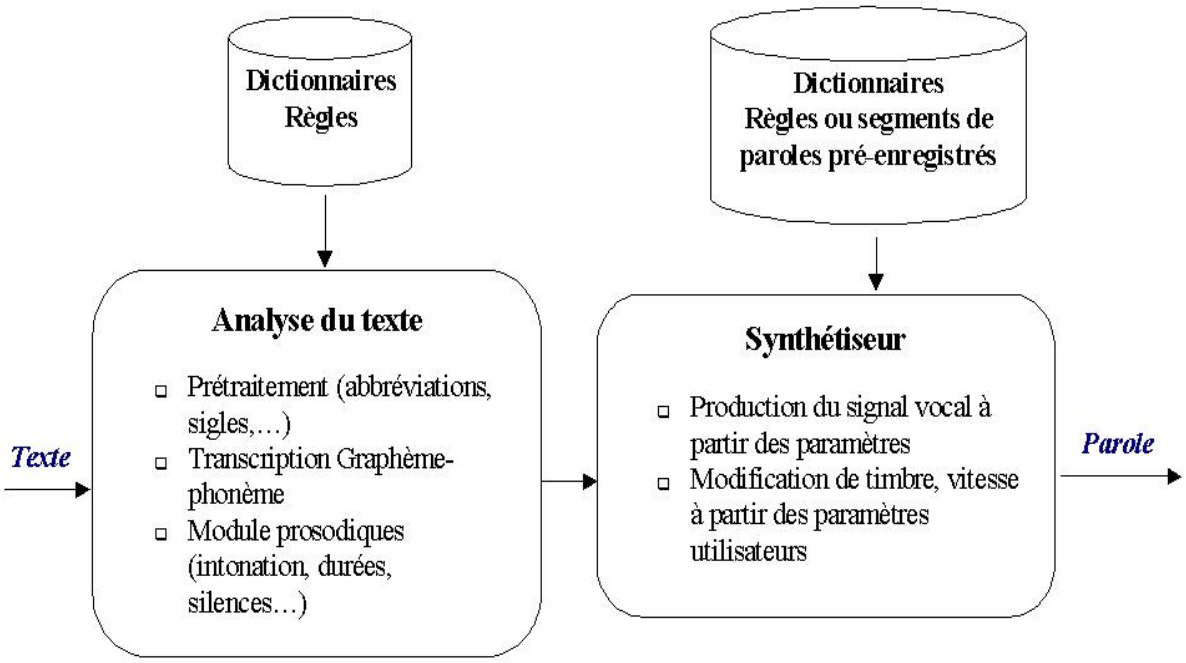


FIG. V.1 – *Architecture classique d'un système TTS*

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

V.3 L'analyse du texte

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

V.3.1 Le prétraitement du texte

Ce module de prétraitement a généralement deux rôles principaux. Le premier est un rôle d'interface entre le texte (représentation linéaire) et la structure de donnés internes gérée par le synthétiseur ([27]). Le second est un rôle d'identification des séquences de caractères qui risquent de poser un problème de prononciation. Parmi les problèmes principalement rencontrés, on peut citer:

– **La détection de fin de phrase** (localisation du point). Ce problème peut s'avérer délicat car le point peut être utilisé dans les nombres (nombres rationnels 3.14), dans les dates (3.3.2001), dans les abbréviations (resp.), dans les acronymes (E.N.S.T), ...
– **Le traitement des abbréviations**. Notons qu'il n'existe pas toujours une transcription unique pour une abbréviation donnée comme on peut le voir sur l'example suivant ([27])): *"Dr. Jones lives at the corner of Jones Dr. And St. James St."*

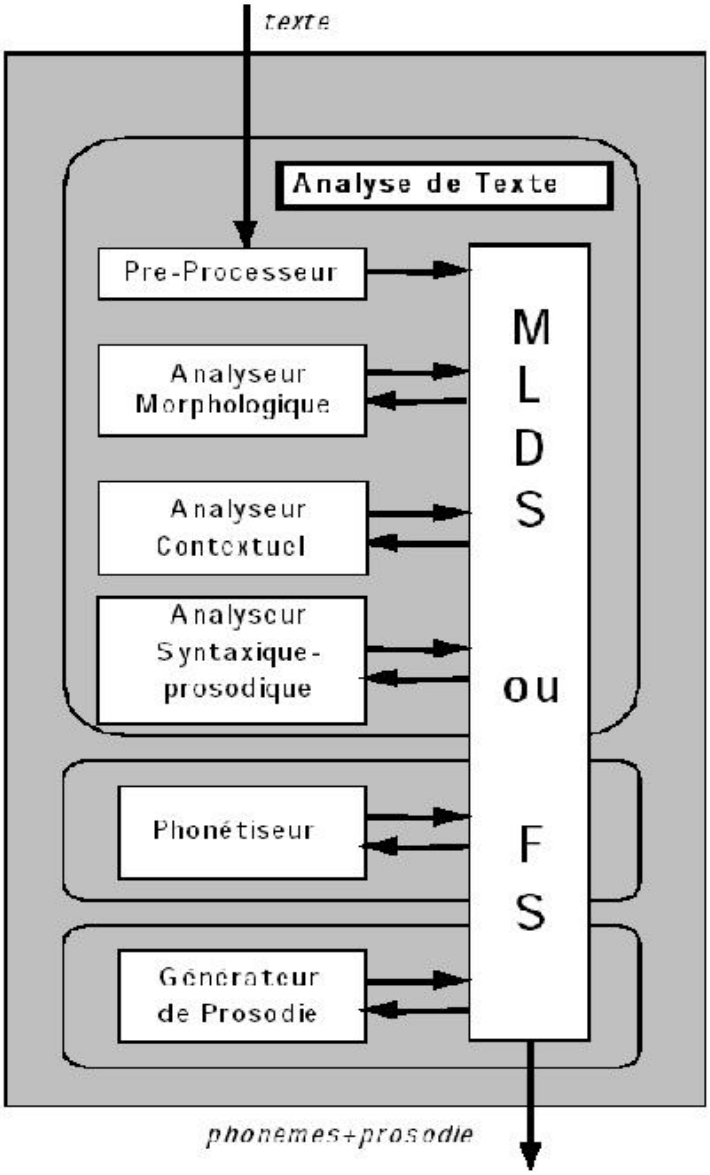


FIG. V.2 – *Module de traitement de texte (d'après [15])*

- **Le traitement des acronymes** et notamment identifier si l'acronyme se prononce (comme pour OVNI ou CNET) ou s'épelle (comme pour SNCF). Cette décision n'est toujours immédiate car pourquoi prononce-t-on CNET et pas ENST?
- **Le traitement des nombres** est un problème à part entière. Il s'agit notamment de pouvoir désambiguiser les nombres rationnels 2.05 des dates 2.05 (2 mai), des heures 2.05 (2h et 5 minutes), d'interpréter les chiffres romains (Henri IV) etc.

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

V.3.2 Analyse morphologique

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

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

Cette analyse distingue en général deux catégories de mots:

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

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

Analyse inflexionnelle

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

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

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

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

ce qui donne les formes suivantes:

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

Analyse dérivationnelle et compositionnelle

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

L'analyse compositionnelle est en fait l'étude de la composition de mots à partir de plusieurs morphèmes de base (par exemple Porte + avion = porte-avion) . Dans certaines langues (comme l'allemand), cette étude peut s'avérer particulièrement complexe (voir [15] pour un exemple). En français, ce problème est toutefois moins délicat.

V.3.3 Analyse contextuelle et syntaxique

Un analyseur contextuel, qui considère les mots dans leur contexte, à pour but

- de réduire le nombre de catégories possibles pour chaque mot en fonction de ses voisins
- de fournir un étiquetage de chacune des unités lexicales constituant la phrase.

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

Il existe un grand nombre d'approches pour l'analyse syntaxique contextuelle. On pourra consulter [15] pour obtenir un premier panorama de ces méthodes. De façon plus succincte, notons qu'il y a 2 principaux types d'approches:

- *Les analyseurs déterministes* qui exploitent un ensemble de règles catégoriques (par exemple de type oui/non). Cette approche est en général assez complexe et nécessite une grande expertise de la langue.

– *Les analyseurs probabilistes* qui utilisent les probabilités de transitions entre catégories syntaxiques successives. Ces approches, assez simples, sont en général très performantes mais nécessitent d'avoir d'importants corpus de données. Nous donnons un exemple ci-dessous à travers les modèles n-grammes.

Etiquetage par n-grammes

Les n-grammes sont utilisés pour estimer la probablité d'une suite de mots w_1,w_2,...,w_n pour un langage donné (notons ici que langage peut représenter une langue mais aussi plus précisément l'utilisation d'une langue dans un domaine d'application particulier). Les modèles les plus répandus sont les *modèles bigrammes* (la probabilité d'un mot ne dépend que de celle du mot précédent) et les *modèles trigrammes* (la probabilité d'un mot ne dépend que de celle des 2 mots précédents). Ainsi, on peut écrire le modèle trigramme d'une suite de mot w_1,w_2,...,w_n sous la forme:

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

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

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}} P(\mathbf{T}|\mathbf{W}) \tag{V.2}$$

En appliquant la règle de Bayes, on obtient:

$$\hat{\mathbf{T}} = \arg\max_{\mathbf{T}} \frac{P(\mathbf{T}|\mathbf{W})}{P(\mathbf{W})} = arg\max_{\mathbf{T}} \frac{P(\mathbf{W}|\mathbf{T})P(\mathbf{T})}{P(\mathbf{W})} \tag{V.3}$$

Pour effectuer l'étiquetage de la suite de mots, on est amené a de plus supposer que:

– La probabilité d'un mot étant donné le passé ne dépend que de son étiquette (hypothèse restreignant l'application du modèle à l'étiquetage syntaxique)
– Le probabiité d'une étiquette étant donné le passé ne dépend que des 2 étiquettes précédentes (modèle trigramme).

Nous pouvons alors écrire que:

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

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

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

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

On associe alors un automate probabiliste à l'équation V.6. Par exemple, pour un modèle bigramme, le modèle comporte M états où chaque état q_i est associé à une étiquette donné t^i. Par souci de simplicité, on autorise toutes les transitions possible, et on utilisera ainsi un

automate complètement connecté (encore appelé ergodique). Pour plus de précisions, on pourra se rapporter à [15].

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

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

V.3.4 Transcription graphème-phonème

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

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

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

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

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

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

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

Les liaisons: elles se caractérisent par l'inclusion d'un phonème à la frontière de deux mots. Ce phénomène est un cas particulier du français. Le nombre de liaisons effectuées dépend du

niveau de langue et du style de prononciation. Certaines liaisons sont obligatoires et sont donc toujours faites (par exemple *'Très utiles'*), dautres sont optionnelles (par exemple *'Deux à deux'*), dautres encore sont interdites (par exemple *'Plat exquis'*). La présence d'une liaison dépend souvent des classes syntaxiques des mots concernés.

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

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

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

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

L'approche par dictionnaire: où le maximum de connaissances morphologiques sont concentrées dans un lexique. Parfois, on utilise des règles morphologiques pour déduire à partir des racines morphologiques stockées dans le lexique, les formes fléchies par dérivation, inflexion ou composition. Dans cette approche, seuls les mots non phonétisés par le dictionnaire sont alors transcrits par règles. C'est l'approche traditionnellement suivie pour langlais américain (MITALK) ([3])

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

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

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

- Le mot "oiseau" se transcrit phonétiquement "/wazo/" , par application des règles suivantes:
  1. la chaîne de caractères orthographiques "oi" se transcrit par la succession des phonèmes /wa/, parce qu'elle est précédée d'un séparateur de mot et qu'elle n'est pas suivie de de la chaîne "gn" comme dans "oignon", ou d'un "n" comme dans "oindre".

2. La lettre "s" se transcrit par le phonème /z/ car cette lettre est entourée par deux voyelles et que "oiseau" ne fait pas partie d'une liste d'exceptions à cette règle, stockée dans le lexique (on pense en particulier à "paraSol" ou "vraiSemblance").
3. La chaîne de caractères "eau" se transcrit par le phonème /o/, indépendamment du contexte.

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

V.4 Génération de la prosodie

V.4.1 Introduction

On oppose souvent le domaine segmental et le domaine prosodique (ou suprasegmental). Pour le domaine segmental, les unités abstraites que sont les phonèmes vont être caractérisées sur le plan acoustique par un ensemble de traits ainsi que l'évolution au cours du temps des formants (i.e. leur fréquence centrale et l'amplitude). Pour le domaine prosodique, ce sont principalement les caractéristiques suprasegmentales (liées aux syllabes ou groupe de syllabes) qui seront concernées.

L'organisation prosodique s'articule donc autour des structures syntaxique, sémantique (relative au sens de l'énoncé) et pragmatique (qui regroupe les informations relatives au contexte particulier de production des actes de parole). Par exemple, il est possible de faire ressortir une syllabe par accentuation (ou focus) et ainsi marquer le mot ou le groupe de mot qui le contient. Ainsi, l'exemple "*un gateau de beurre Suisse frais*" peut donner lieu à plusieurs interprétations différentes en fonction de la position des accents.

Il existe enfin des contraintes extra-linguistiques qui influent sur les paramètres prosodiques : les contraintes phonotactiques d'une part qui sont relatives au nombre de syllabes par mot, par groupe syntaxique et par phrase. Enfin, le contexte d'élocution, l'émotivité du sujet, la constitution physiologique des organes de production (liés principalement au sexe et à l'âge du locuteur), l'origine régionale et sociale du locuteur seront autant de paramètres qu'il faudra prendre en compte dans la structure prosodique de l'énoncé([10]).

En synthèse de la parole, la prosodie joue un rôle primordial dans le naturel et l'intelligibilité de la parole synthétique, et il est donc important de la modéliser avec soin.

La prosodie est caractérisée par trois paramètres prosodiques principaux. Il s'agit de:

La fréquence fondamentale: qui représente l'estimation de la fréquence laryngienne à partir du signal acoustique à un instant donné.

La durée: La durée est une notion plus large. Elle inclut le débit de parole, la durée et la répartition des pauses, les allongements syllabiques .... Elle correspond cependant toujours à la mesure d'un intervalle de temps donné.

L'intensité: qui correspond à l'énergie contenue dans le signal au cours d'un intervalle de temps donné.

Il existe différentes manières de définir les paramètres prosodiques, selon qu'on les considère sur le plan de la production, sur le plan acoustique, ou perceptif. La figure V.3 résume les différents termes utilisés selon l'endroit où l'on se place dans la chaîne de production-réception.

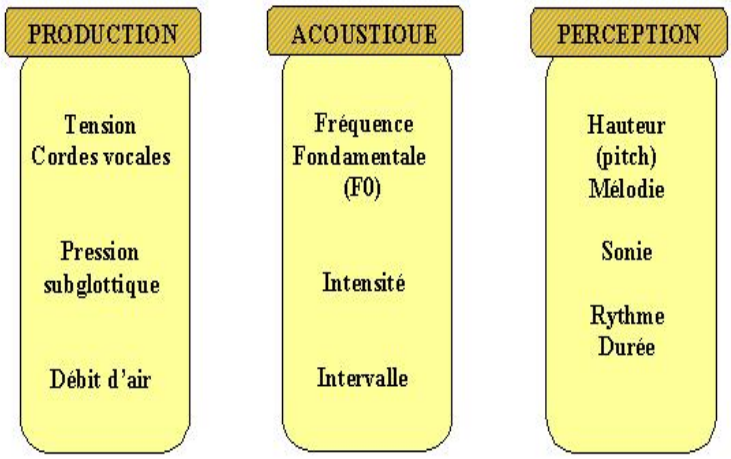


FIG. V.3 – *Les différentes définitions des paramètres prosodiques (d'après [10])*

Lorsque l'on se situe d'un point de vue acoustique, l'étude des paramètres se base sur ce qui est mesurable. Les paramètres peuvent être alors considérés comme parfaitement indépendants. Cependant, sur le plan de la production ou de la perception, il existe une forte interaction entre les différents paramètres.

V.4.2 La mélodie

L'étude de la mélodie est souvent séparée en deux types de phénomènes:

les phénomènes micromélodiques: qui sont relatifs à des contraintes physiologiques et/ou acoustiques sur l'appareil phonatoire lors de l'articulation de certains types de phonèmes. Bien qu'elle peut jouer un rôle dans l'identification des phonèmes, son influence reste mineure et n'est en général pas modélisée dans les systèmes de génération automatique de la prosodie.

les phénomènes macromélodiques: qui concernent les événements principaux de la prosodie. Ces phénomènes peuvent eux-mêmes se diviser en deux classes distinctes :

- *Les événements de portée locale:* qui tiennent compte de tous les événements locaux qui sont relatifs à la mise en relief d'une syllabe ou d'un mot. Ces événements locaux constituent un des indices acoustiques majeurs de la réalisation de l'accent. Cette entité, définie comme la forme mélodique comprise entre deux minima locaux (voir F1, F2, F3, F4 figure V.4, est reliée à la notion de "hat-pattern" (t'Hart et al., 91).
- *Les événements de portée globale:* qui concernent les événements mélodiques qui s'étendent sur des portions de parole beaucoup plus longues, voire même sur la totalité de la phrase. Le phénomène mélodique global le plus important concerne la tendance que connaît la fréquence fondamentale à décroître lentement du début à la fin de la phrase (ligne de déclinaison).
  La déclinaison est souvent accompagnée du phénomène de "remise à zéro" ("resetting"), notamment lorsque l'on considère des phrases longues. Ces réinitialisations de la déclinaison seront généralement situées à des frontières de groupes syntaxiques, accompagnées la plupart du temps d'une pause.

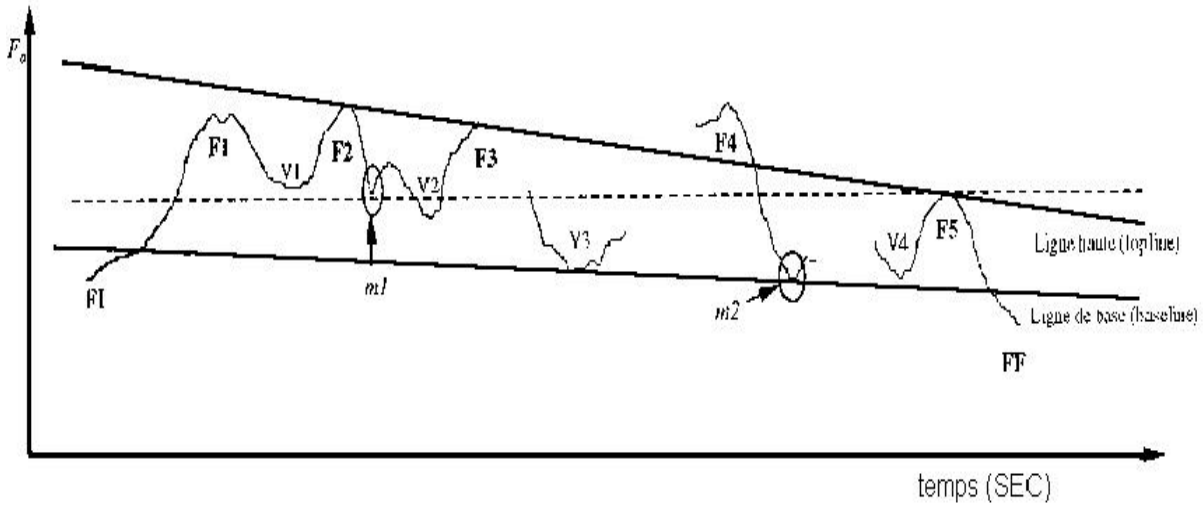


FIG. V.4 – *Présentation des principaux paramètres permettant de caractériser les événements mélodiques présents lors d'une analyse acoustique ; FI = fréquence initiale, FF = fréquence finale, Vx = vallées, Fx =pics mélodiques, mx = creux micromélodiques (d'après [9])*

V.4.3 La durée

Des trois paramètres prosodiques, la durée est le plus difficile à préciser, car elle n'est pas directement associable à un corrélat biologique du système phonatoire ([10]). On mesurera un intervalle de temps relatif à une unité choisie, qui pourra être la durée du phonème. L'étude du paramètre de durée dépend de la segmentation et de l'étiquetage.

La durée des phonèmes est très dépendante du contexte dans lequel ils apparaissent. D'autre part, le découpage de l'énoncé en groupes prosodiques va se traduire sur le plan acoustique sous la forme d'allongements vocaliques et de pauses. Ces deux composantes de variations de durée sont mêlées et occasionnent des distorsions très importantes de l'axe temporel. Lors de la mise en place d'une étude sur le rythme de la parole, il sera donc primordial d'appliquer une méthode permettant la séparation des deux composantes. Plusieurs méthodes de "normalisation" des durées ont été proposées.

V.4.4 L'intensité

L'intensité est liée à l'ouverture glottique et à la pression d'air subglottique. Lors de l'analyse des paramètres prosodiques, on entend par intensité, l'énergie contenue dans le signal de parole durant un intervalle de temps donné.

V.4.5 Segmentation en groupes prosodiques

Comme il est dit précédemment, la prosodie utilise plusieurs niveaux de connaissance incluant la sémantique (le sens de la phrase) qui n'est en général pas accessible au synthétiseur. Ainsi, l'une des fonctions les plus importantes de la prosodie qui pourra être réalisée par un synthétiseur sera celle de la segmentation d'un énoncé en groupes prosodiques (qui sont ici des groupes de mots plus petits et qui permettent de rendre un énoncé plus clair). Il existe plusieurs types de

segmentation incluant:

La segmentation heuristique: qui est basée sur l'établissement de règles simples basées sur la ponctuation (par exemple pour la phrase *"Le gateau, que j'ai mangé, était excellent."* une séparation sera marqué au niveau de chaque virgule et point). Des améliorations peuvent être apportées en tenant compte de la distinction mot lexical/mot grammatical. Dans l'exemple "la *table* de mon *oncle* est *noire*", nous aurons une séparation après chaque mot lexical (mots en italique)

La segmentation par analyse morphosyntaxique Cette approche est basée sur le fait qu'il existe une certaine congruence entre la syntaxe et la prosodie. Cette congruence étant approximative il est souvent nécessaire d'ajouter des règles d'ajustement incluant par exemple le principe d'eurythmie. Dans l'exemple *"Le père de Marie est venue"*, les différentes segmentations possibles (en nombres de syllabes) font apparaître soit une segmentation en 2 groupes principaux (2+3)(2) ou (2)(3+2) soit en trois groupes de taille équivalente (2)(3)(2). Le principe d'eurythmie favorisera un découpage en groupes de taille équivalente et donc le dernier exemple donné ci-dessus.

Segmentation par apprentissage: Dans cette approche, les règles sont apprises automatiquement à partir de grands corpus de texte préalablement étiquetés. L'une des approches les plus répandues utilise des arbres de décisions et des techniques de classification et régression (l'approche CART: Classification and Regression Tree). Cette approche permet de mettre automatiquement en évidence les facteurs contextuels les plus significatifs. Les CART fonctionnent comme des arbres binaires. A chaque noeud une décision est prise et permet de séparer les données en deux groupes et ainsi de suite. Les CART travaillent aussi bien sur des données symboliques (décision) que des données numériques (ici c'est l'intervalle de variation du paramètre en question qui est séparé en deux). On trouvera un exposé plus détaillé dans [15] ainsi que de nombreuses références. Notons toutefois que cette approche a donné des résultats très satisfaisant pour également prédire la durée des phonèmes ou la position des pauses. Cette approche est celle suivie par plusieurs systèmes de synthèse incluant le synthétiseur FESTIVAL et celui des Bell Labs ([90],[13]).

V.4.6 Les modèles de génération de la mélodie

Dans le cadre de la génération automatique de la prosodie en synthèse de la parole, les trois paramètres prosodiques sont généralement générés indépendamment. Nous abordons successivement ci-dessous plusieurs méthodes pour la génération de l'intonation et présentons brièvement les approches pour la génération des variations de durée.

À l'heure actuelle, on peut dénombrer cinq principales méthodes de modélisation, chacune d'entre elles ayant déjà été appliquée à différentes langues ([10]):

le modèle de commande de source vocale

Cette méthode suppose que la courbe mélodique peut être construite à l'aide de deux types de commandes (des commandes d'accent et des commandes de groupe). Ces commandes prennent leur origine dans la physiologie et visent à rendre compte de l'activité musculaire laryngienne mise en jeu lors de la production de la mélodie.

Ce modèle suppose ainsi que l'information prosodique (structures et unités) programmée par le locuteur est fondée sur des événements discrets, alors que sa réalisation acoustique est continue en temps et en fréquence. Ceci serait une simple conséquence liée aux mécanismes physiologiques impliqués dans le processus de contrôle de la fréquence fondamentale. Il existe donc une fonction

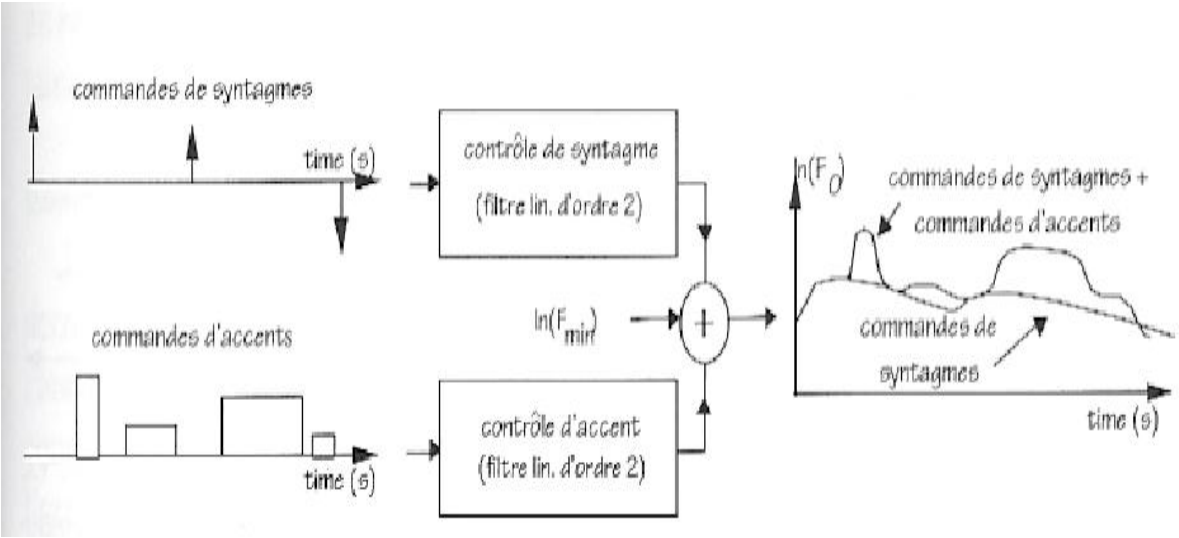


FIG. V.5 – *Modèle fonctionnel de commande de source vocale de H. Fujisaki (d'après [37])*

de transfert (liée à des contraintes physiologiques) reliant les commandes ponctuelles à la courbe mélodique réellement produite. Le modèle de Fujisaki propose ainsi de décomposer la courbe mélodique en deux types de commandes:

- *Les composantes de phrases* qui correspondent à la réponse d'un système linéaire du second ordre à des impulsions,
- *les composantes d'accents* qui correspondent à la réponse d'un autre système linéaire du second ordre à des "fonctions échelon"

La figure V.5 montre le schéma général du modèle de Fujisaki où les commandes des deux types sont superposés pour former un contour mélodique.

la méthode par points cibles (ou approche tonale)

Cette méthode présuppose que l'information mélodique est principalement contenue dans les extrema de la courbe. Ce modèle est d'inspiration phonologique car il s'agit de simplifier la courbe intonative par une suite de tons discrets (points cibles) reliés entre eux par des fonctions de transition.

Les cibles sont définies à l'intérieur d'une enveloppe qui détermine la dynamique globale. Les transitions entre les cibles sont définies par règles.

Une fois la courbe mélodique totalement étiquetée par rapport à la structure segmentale à l'aide des "étiquettes tonales", plusieurs étapes sont nécessaires pour générer la réalisation acoustique de cette mélodie. Par exemple dans le système de Pierrehumbert ([74]), on définit dans un premier temps le registre (pitch range). On en déduit la dynamique, qui est formée d'une ligne de base et d'une ligne haute calculées indépendamment. Ces lignes forment "l'enveloppe" du contour mélodique. Le contour sera ensuite exprimé comme une fraction de la distance entre ces deux lignes.

Chez Pierrehumbert, deux niveaux sont utilisés pour les *points cibles*: le point H (représentant un ton haut) et le point L (représentant un ton bas). Des *fonctions de transitions* sont ensuite utilisées pour relier les tons successifs (par exemple, deux H consécutifs sont interpolés par une

fonction de transition qui abaisse la fréquence entre les deux cibles). Notons qu'il est possible de choisir des fonctions de transitions variées sachant qu'elles ne jouent pas un rôle prépondérant dans la mélodie. Notons également qu'il est possible de proposer un plus grand nombres de tons (Mertens en a par exemple proposé 4 : H pour, L pour bas, pui H+ et L+ pour respectivement extra-haut et infra-bas pour rendre compte de saut de pitch plus important [66]).

Notons enfin, que certaines études actuelles visent, à partir de cette approche, à obtenir de façon automatique des règles permettant de générer automatiquement les contours mélodiques. Ces règles sont alors inférées à partir de grands corpus de données.

l'école hollandaise:

La technique développée consiste à simplifier la courbe mélodique globale en une concaténation de segments de droites (selon une échelle logarithmique). Cette simplification est réalisée sur la base de tolérances perceptives à l'aide de la technique d'analyse-synthèse.

La première étape de simplification consistera à obtenir des contours réduits à un nombre minimum de segments linéaires, qui soient indiscernables des contours naturels: c'est l'étape de *stylisation*. La seconde étape correspondra à la classification des mouvements obtenus en un nombre restreint de classes standard, de manière à fournir une description des propriétés acoustiques des unités mélodiques de base pour la langue étudiée: c'est l'étape de *standardisation*.

Cette approche a aussi été adoptée pour certains systèmes français ([9]). Dans un tel système, les étapes suivantes sont considérées:

- *Découpage en groupes prosodiques*. Ce découpage est réalisé en fonction de la structure syntaxique, de la catégorie des mots, et de critères rythmiques (longueur de la phrase, longueur des groupes en nombre de syllabes). Il permettra de positionner quatre types de frontière entre ces groupes (ponctuation, majeure, mineure, terminale).
- *Génération des accents de groupes :* pour chaque groupe, un accent initial et un accent final peuvent potentiellement être réalisés. La validation de ces accents dépendra principalement de la longueur des groupes et de la contiguïté éventuelle de ces accents.
- *Réalisation phonétique des accents:* La réalisation phonétique (et acoustique) des accents est alors déduite de la structure accentuelle schématisée. C'est ici que sera réalisé le lien entre les marqueurs accentuels et la description mélodique réalisée en mouvements standards.

La figure V.4.6 nous donne un exemple de génération automatique à l'aide de ce système.

La modélisation sous forme de contours mélodiques stockés

Une autre approche assez populaire pour la modélisation de la prosodie vise à modéliser la fréquence fondamentale de la parole en contours prototypiques. Le stockage des unités de référence est réalisé en fonction d'hypothèses fortes sur la manière dont les informations linguistiques sont véhiculées.

Notons toutefois, que ces systèmes visent à découper une phrase en groupes prosodiques généralement liés à la syntaxe à l'aide de règles plus ou moins complexe puis à ensuite représenter l'intonation à l'aide de contours types enregistrés dans un lexique.

Notons également qu'il est possible d'obtenir ces contours par des approches statistiques en utilisant par exemple un réseau de neurones avec en entrée les marqueurs de phrase, syntagme et groupe prosodique.

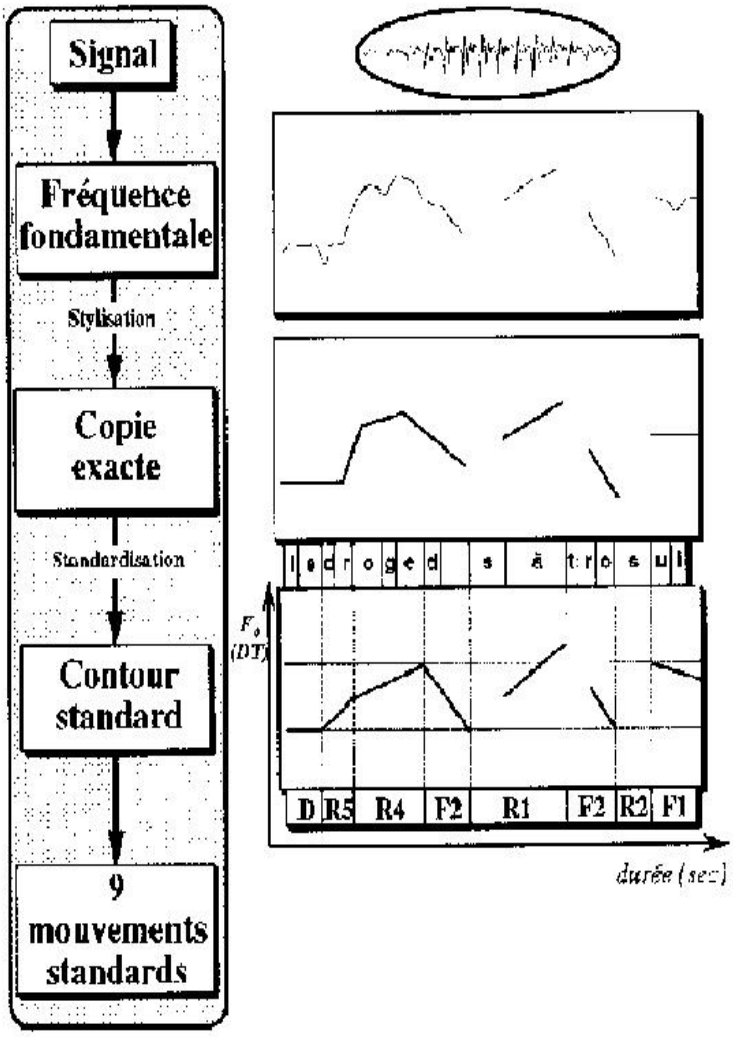


FIG. V.6 – *Schéma général de la méthodologie de l'école hollandaise appliquée au français(d'après [10])*

Génération de contours intonatifs à partir de systèmes statistiques

Plus récemment de nouvelles approches utilisant de larges corpus étiquetés se sont développées. Il s'agit, à l'aide de méthodes statistiques ou de réseaux neuronaux, d'associer description linguistique et substance prosodique. Ces modèles nécessitent d'utiliser une caractérisation linguistique appropriée et de prélever des valeurs de substance prosodique qui permettent de rendre compte le mieux possible des phénomènes prosodiques pertinents.

C'est surtout dans la préparation des paramètres de caractérisation que réside la plus grande partie du travail. Si ces paramètres sont bien choisis, on peut alors arriver à générer des contours mélodiques de très bonne qualité sans pour autant savoir comment se fait le passage entre les niveaux de représentation "supérieurs" et la substance elle-même.

Ces systèmes fournissent l'avantage important d'être capables de s'adapter à de nouvelles bases de données automatiquement et très rapidement à partir du moment ou les niveaux de description restent valides (voir par exemple [94],[63])

V.4.7 Modèles de durée

Les modèles de durée utilisent des sources d'informations diverses incluant:

- Les informations de phonéticiens (basées sur des principes articulatoires ou phonologiques)
- L'estimation de paramètres basée sur des expériences où un petit nombre de paramètres varient. Ces expériences sont généralement menées sur des bases de données de taille restreinte autorisant une analyse manuelle détaillée.
- L'estimation de paramètres basée sur des expériences de grande ampleur (utilisation de corpus de grande taille)

Une approche couramment retenue consiste à estimer les durées intrinsèques (i.e. leurs valeurs moyennes sur un grand corpus) puis à modifier ces durées en faisant intervenir des facteurs co-intrinsèques (durées des phonèmes voisins) et linguistiques sous la forme de facteurs multiplicatifs ou additifs.

La disponibilités de gros corpus de données étiquetés permet de plus en plus d'évoluer vers des modèles plus généralistes qui peuvent contrôler un grand nombre de paramètres. Les approches utilisées sont du même types que les méthodes statistiques utilisées pour l'analyse syntaxique ou les modèles de génération de la mélodie (i.e. CART, réseaux de neurones,) ([19][63]).

V.5 Synthèse de la parole

L'étape précédente a permis de passer d'une information graphémique (le texte) en une information phonétique (les phonèmes à prononcer) associée à une information prosodique (l'intonation à reproduire). L'étape de synthèse a proprement parlé consistera à générer le signal de parole à partir des informations pré-citées. Il existe de nombreuses approches pour réaliser cette synthèse. Ces approches sont classiquement ordonnées en fonction de leur niveau de description allant des modèles fonctionnels (synthèse par concaténation d'éléments sonores pré-enregistrés) aux modèles physiques (synthèse articulatoire). La figure V.7 donne une telle classification des techniques de synthèse en précisant les caractéristiques principales de chaque approche.

Nous avons vu au chapitre III les approches par modélisation articulatoire. Ce chapitre sera ainsi consacré aux deux autres approches classiques que sont la synthèse par concaténation et la synthèse par règles (ou synthèse à formants).

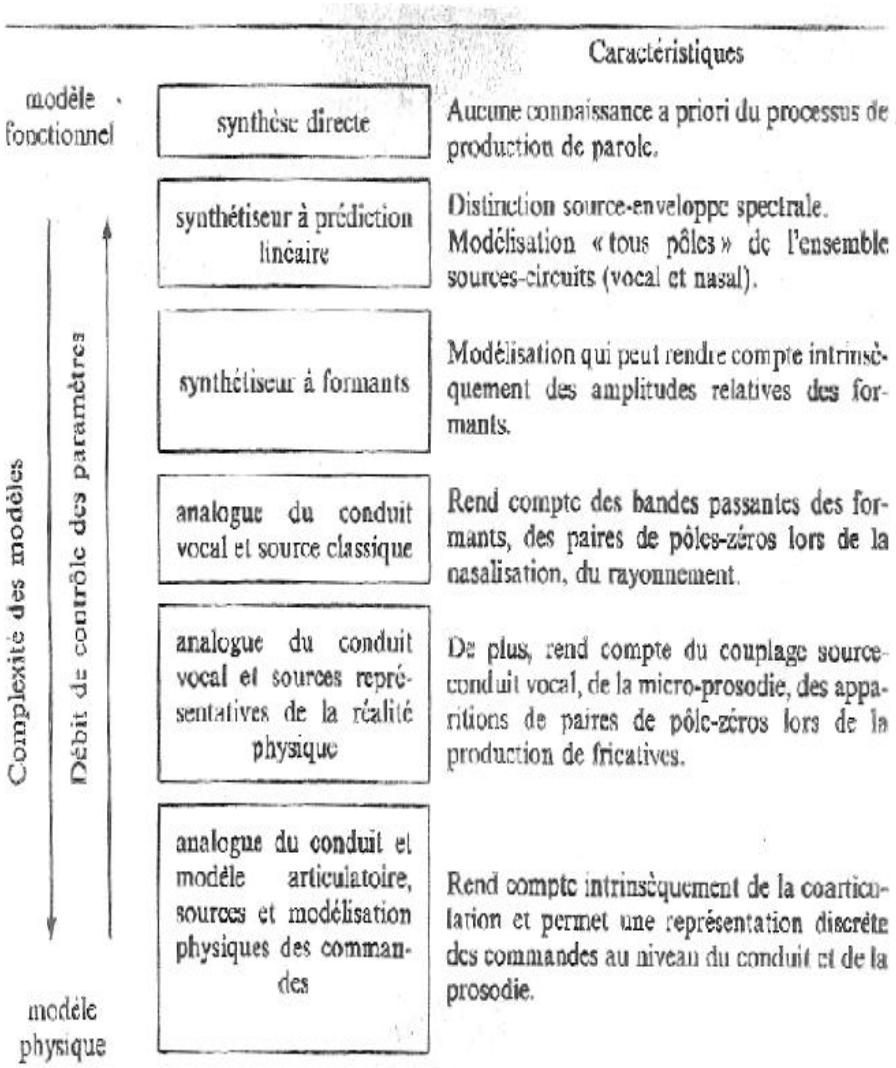


FIG. V.7 – *classification des techniques de synthèse (d'après [18])*

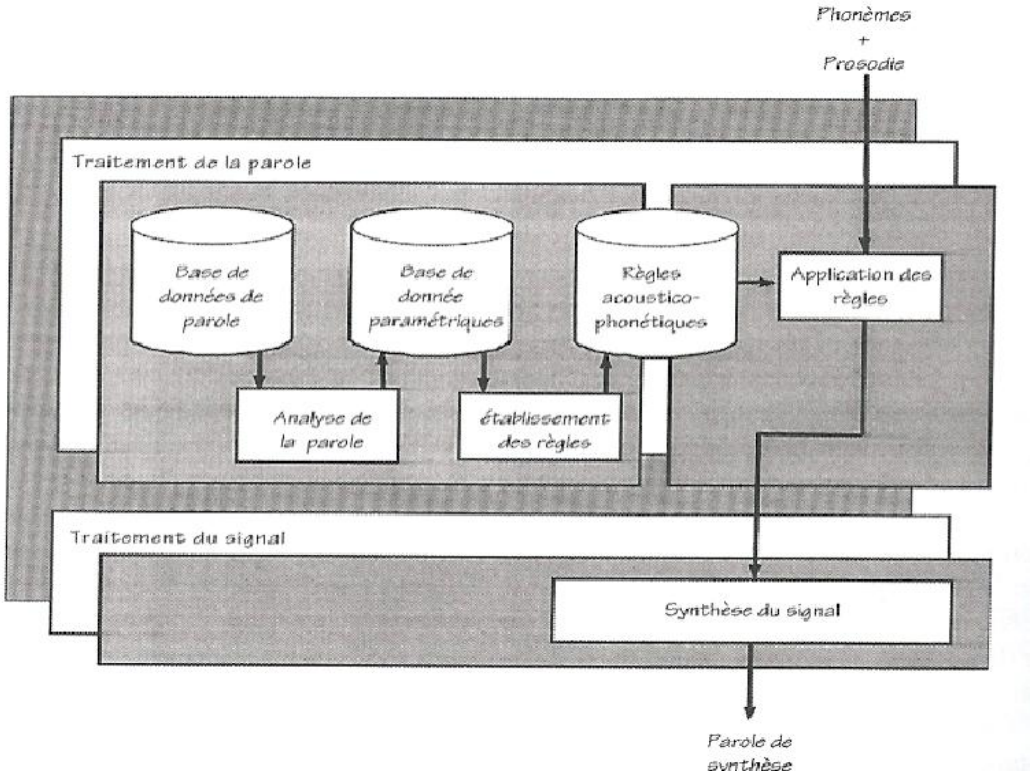


FIG. V.8 – *Schéma de conception et fonctionnement typique d'un système de synthèse par règles(d'après [15])*

V.5.1 Synthèse par règles

Les synthétiseurs par règles sont les premiers systèmes qui ont donné de la synthèse de bonne qualité ([53, 54]). Ils sont basés sur l'idée que, si un phonéticien expérimenté est capable de "lire" un spectrogramme, il doit lui être possible de produire des règles permettant de créer un spectrogramme artificiel pour une suite de phonèmes donnés. Une fois le spectrogramme obtenu, il ne reste plus alors qu'a générer l'audiogramme correspondant. La figure V.8 donne un schéma général d'un synthétiseur par règles.

La mise au point d'un tel système de synthèse suit les étapes suivantes:

1. Dans un premier temps, on fait lire par un locuteur professionnel un grand nombre de mots, généralement de type Consonne-Voyelle-Consonne (CVC) et on les enregistre sous forme numérique. Les mots sont choisis de façon à constituer un corpus représentatif des transitions phonétiques et des phénomènes de coarticulation dont on veut rendre compte.
2. Dans un second temps, on modélise alors ces données numériques à l'aide d'un modèle paramétrique de parole, qui a pour rôle de séparer les contributions respectives de la source glottique et du conduit vocal et de présenter cette dernière sous forme compacte (par exemple visualisation des formants), plus propice à l'établissement des règles.
3. Ensuite, les règles sont établies. On précise alors les valeurs numériques des paramètres intervenant dans ces règles (les fréquences des formants, ou les durées des transitions, par exemple) par un examen minutieux du corpus. Notons que cette étape d'estimation est

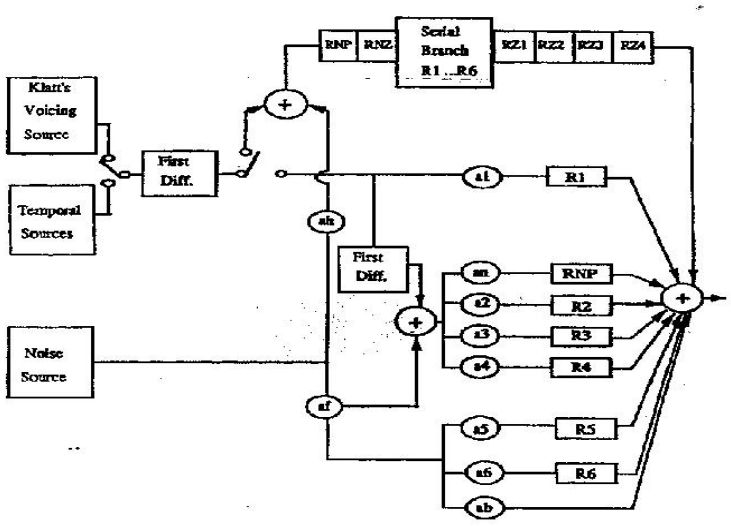


FIG. V.9 – *Architecture Série-parallèle pour un synthétiseur à formants (d'après [54])*

menée sur une seule voix : un moyennage inter-locuteur aurait peu de signification dans ce contexte. La mise au point du synthétiseur s'achève par un long processus d'essais-erreurs, afin d'optimiser la qualité de la synthèse.

4. Lorsqu'un nombre suffisant de règles ont été établies, la synthèse proprement dite peut commencer. Les entrées phonétiques du synthétiseur déclenchent l'application de règles, qui produisent elles-mêmes un flux de paramètres liés au modèle de parole utilisé. Cette séquence temporelle de paramètres est alors transformée en parole par un synthétiseur, qui implémente les équations du modèle. Une architecture classique pour un tel synthétiseur est celle proposée par Klatt (voir figure V.9 ([54]):

La synthèse par règles possède plusieurs avantages et notamment celui de nécessiter très peu de mémoire pour stocker les données. Elle possède cependant de nombreux inconvénients. En effet, l'établissement des règles est long et fastidieux. De plus, ces règles dépendent en grande partie de la langue ce qui nécessite un travail important pour l'établissement d'une nouvelle langue de synthèse. Notons également que les règles ont été construites à partir de la voix d'un locuteur et sont ainsi dépendantes de ce dernier. Finalement, malgré cette complexité de mise au point, les synthétiseurs à formants ne parviennent pas à produire une voix de synthèse ayant la même qualité que les synthétiseurs par concaténation d'éléments sonores. Ainsi, si la synthèse par règles a connu un essor considérable dans les années 60-70, elle n'est plus guère utilisée aujourd'hui que lorsque les contraintes de mémoire et de temps de calcul sont très importantes.

V.5.2 Synthèse par concaténation

Au contraire des synthétiseurs par règles, les synthétiseurs par concaténation ont une connaissance très limitée du signal qu'ils mettent en forme. La plupart de ces connaissances se trouve en effet stockée dans les unités de parole mises en oeuvre par le synthétiseur. Ceci apparaît clairement dans la description générale d'un tel synthétiseur (voir figure V.10), où l'on constate que la plupart des opérations liées à la synthèse proprement dite (par opposition aux opérations nécessaires à la création du synthétiseur) se retrouvent groupées dans un bloc de traitement du signal ne faisant aucune référence explicite à la nature profonde des signaux traités. La synthèse

par concaténation procède en effet par mise bout à bout de segments acoustiques déjà coarticulés, extraits d'une base de données de signaux de parole. Il s'ensuit que, contrairement aux cibles phonétiques de l'approche précédente, qui nécessitent l'établissement de règles (phonétiques) pour modéliser correctement leurs transitions, la production de parole fluide en synthèse par concaténation ne requiert qu'une étape de concaténation qui s'accompagne d'un lissage purement acoustique des discontinuités pouvant apparaître aux points de concaténation. Comme pour la synthèse par règles, un certain nombre d'opérations préliminaires doivent être menées avant que le synthétiseur ne soit capable de produire sa première parole. C'est le rôle des modules de traitement de la parole.

Sélection des unités de synthèse

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

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

Constitution de la base de données de segments

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

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

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

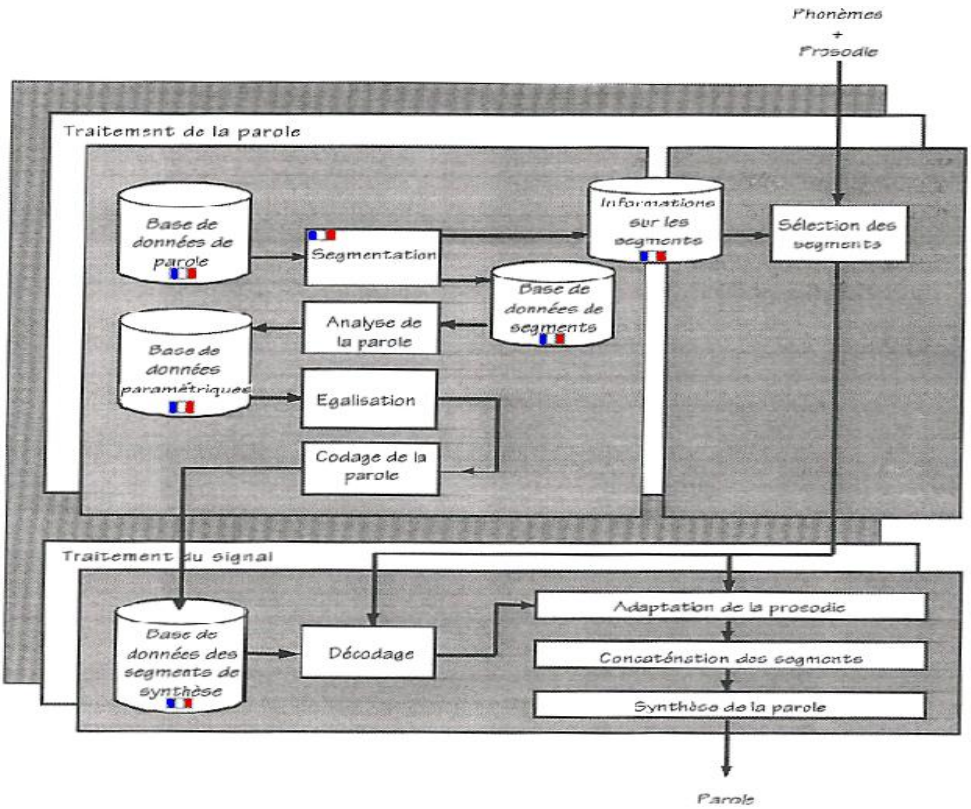


FIG. V.10 – *Schéma général d'un synthétiseur par concaténation. Les opérations qui dépendent de la langue sont indiquées par un drapeau.(d'après [15])*

utilisables. On centralise également l'information relative à ces segments (leur nom, leur durée, leur pitch, et les marqueurs de frontières de phonèmes à l'intérieur des segments) dans une base de données séparée, qui sera utilisée par le bloc de sélection d'unités. Dans le cas de diphones, par exemple, on mémorise l'instant de passage d'un phonème à l'autre afin de pouvoir plus tard modifier séparément les durées de chaque demi-phonème.

Modélisation paramétrique et codage

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

- Un modèle bien choisi permet souvent une réduction de la taille des données. On pourra donc se permettre de stocker plus d'unités pour une même quantité de mémoire, ou réduire la taille mémoire nécessaire pour un nombre donné d'unités. Ceci justifie la présence d'un codeur de parole sur la figure V.10. Cet avantage est important en synthèse par concaténation étant donné le grand nombre d'unités à stocker.
- De nombreux modèles de parole séparent explicitement les contributions respectives de la source et du conduit vocal . Ceci est mis à profit par le synthétiseur pour résoudre indépendamment (et donc plus simplement)les deux problèmes fondamentaux évoqués plus haut: la modification de la prosodie des unités et leur concaténation.
- De même, certains modèles séparent explicitement la parole en deux contributions (une contribution voisée et une contribution non-voisée) ce qui permet aussi d'améliorer les problèmes de raccordement des unités au moment de la synthèse

Sélection des segments

Lorsque la base de données des segments de synthèse a été constituée, la synthèse proprement dite peut commencer. Les informations phonétiques et prosodiques présentées à l'entrée du synthétiseur sont tout d'abord transformées en séquences de commandes de segments du synthétiseur. Ceci est réalisé à l'aide du module de sélection de segments (ou d'unités) de la figure V.10. La distinction introduite plus haute entre sélection segmentale d'unités et sélection totale (segmentale et suprasegmentale) est bien entendu d'application ici. En sélection segmentale, la suite des unités à concaténer est déduite de la chaîne phonétique d'entrée uniquement. Au contraire, la sélection totale implique un choix d'unités réalisant au mieux les caractéristiques segmentales et suprasegmentales (typiquement: pitch et durée) de la chaîne phonétique d'entrée. Que ce soit en sélection segmentale ou totale (la première n'étant qu'un cas particulier de la seconde), deux cas de figure peuvent se présenter.

Sélection statique Dans le premier, chacune des unités à synthétiser peut être déduite indépendamment des autres, directement à partir de la suite des phonèmes à produire. C'est le cas par exemple d'une synthèse avec sélection segmentale de diphones dans une base de données ne contenant qu'une seule instance de chaque diphone: la détermination de chaque diphone ne dépend que d'un couple de phonèmes successifs dans la chaîne phonétique d'entrée. On parle

alors de sélection statique. Notons que la plupart des systèmes actuels suivent encore cette stratégie.

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

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

Concaténation des segments

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

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

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

Modification de la fréquence fondamentale et de la durée

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

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

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

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

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

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

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

Méthode temporelle TD-PSOLA[3] Cette méthode suppose que l'on traite un signal de parole dont on connaît la période fondamentale. L'idée [68] est encore fondée sur l'hypothèse que le signal de parole est constitué d'impulsions glottales filtrées par le conduit vocal. On observe ainsi une succession de réponses impulsionnelles, positionnées à des temps multiples de la période (hypothèse du peigne temporel convolué avec la réponse impulsionnelle du conduit vocal).
On définit d'abord des 'marques d'analyses' synchrones de la fréquence fondamentale pour les parties voisées, positionnées sur la forme d'onde à chaque période. Les modifications d'échelles sont alors effectuées de la façon suivante:

Modification de l'échelle temporelle Pour modifier la durée du signal sans en altérer la fréquence fondamentale, on va simplement dupliquer (étirement temporel) ou éliminer (compression temporelle) des périodes de la forme d'onde, en fonction du taux de modification désiré. On est donc conduit à définir des marques de synthèse également synchrones du fondamental, associées aux marques d'analyse (de façon non-bijective puisque certaines marques sont dupliquées ou éliminées).
Les signaux à court-terme situés autour de chaque marque d'analyse sont alors extraits (par l'utilisation d'une fenêtre temporelle-par exemple de type hanning- de durée égale à deux périodes et centrée sur la marque d'analyse) et 'recopiés' autour des marques de synthèse correspondantes et le signal modifié est obtenu par une simple méthode d"overlap/add". La figure V.11 illustre le principe de cette méthode pour un taux d'étirement temporel local de 1.5.
On voit que deux périodes du signal original ont donné naissance à trois périodes dans le signal modifié, ce qui correspond bien à un étirement temporel mais la durée de la

3. Paragraphe écrit par J. Laroche [55]

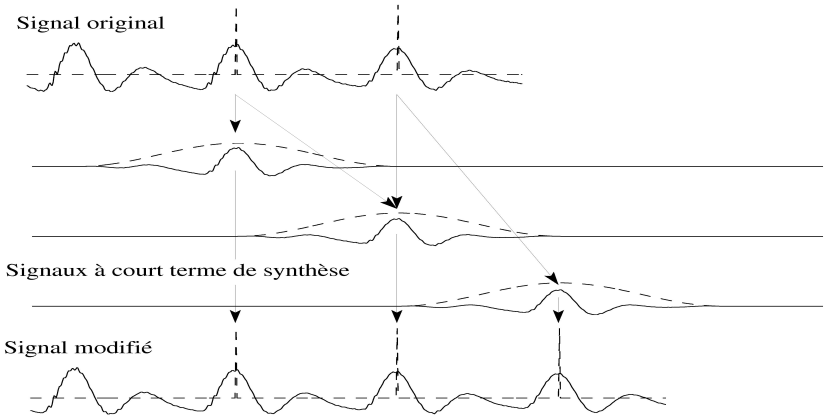


FIG. V.11 – *Modification de la durée du signal par la méthode TD-PSOLA. En haut, le signal original, au milieu trois signaux à court-terme générés à partir des deux signaux à court-terme centrés autour des deux premières marques d'analyse. En bas, signal modifié.*

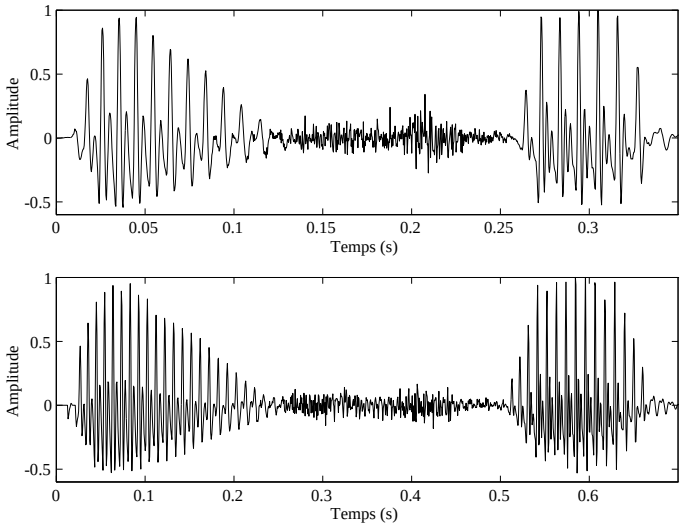


FIG. V.12 – *Original: "il s'est"(d'après [55])*

période n'est pas modifiée (l'écartement des marques de synthèse est le même que celui des marques d'analyse), la fréquence fondamentale du signal est conservée.

La figure V.12 donne un exemple d'application à la phrase 'il s'est' dont l'original est donné en haut de la figure. On remarque la partie non-voisée au centre de la fenêtre (le son 's'), séparant les deux parties voisées /i/ et /e/.

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

On est ainsi conduit à définir des marques de synthèse correspondant à la valeur modifiée du fondamental, et à les associer aux marques d'analyse comme précédemment. Puisque les marques de synthèse sont plus serrées (élévation du fondamental) ou écartées (abaissement du fondamental) que dans le signal original, il faut pour conserver la durée du signal dupliquer ou éliminer certaines marques. La figure V.13 illustre le principe de cette méthode.

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

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

La méthode décrite ci-dessus réalise des modifications de très bonne qualité. Par sa simplicité, elle peut faire l'objet d'une implémentation temps réel. Par contre, il est important de noter

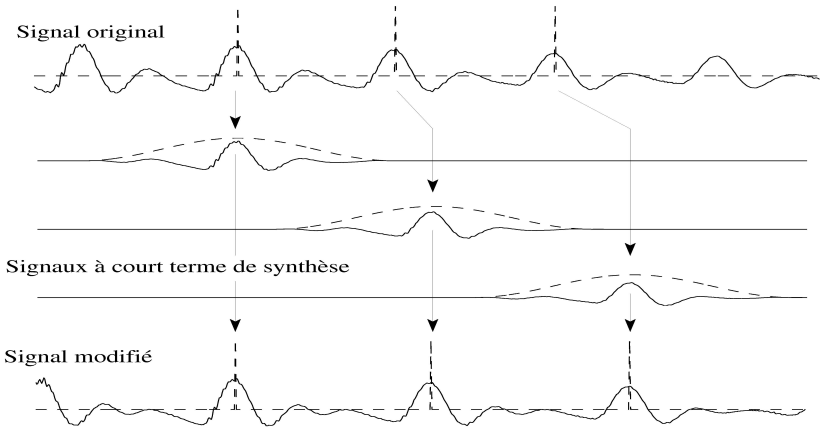


FIG. V.13 – *Modification de la hauteur du signal par la méthode TDPSOLA. En haut, le signal original, au milieu trois signaux à court-terme générés à partir des trois premières marques d'analyse. En bas, signal modifié. L'écartement des marques de synthèse n'est pas identique à celui des marques d'analyse.*

que la qualité des modifications de fondamental sont très sensibles à la position des marques d'analyse. On peut alors se tourner vers d'autres méthodes (notamment fréquentielles). Pour d'autres méthodes basées sur des idées très similaires, on pourra se référer à [69], ou à l'article de J. Laroche dans [52].

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

Lissage des segments

Comme on l'a vu, une approche telle que TD-PSOLA permet de modifier la durée et/ou la fréquence fondamentale d'un segment mais ne dispose pas de moyen simple pour lisser les discontinuités spectrales. Pour éviter les discontinuités spectrales, il est alors possible de soit optimiser la base des segments de telle sorte que ces discontinuités soient réduites au minimum soit utiliser des techniques de lissage plus automatisés. La seconde approche est intéressante car elle est indépendante de la langue et permet d'obtenir plus rapidement une base de segments de très bonne qualité. La plupart des approches vise à re-synthétiser la base de segments pour lui conférer certaines propriétés. Par exemple, l'approche MBROLA resynthétise la base de segments de telle sorte que tout les segments sont à pitch constant et à phase fixe (les marqueurs de période sont placés au même endroit sur chaque segment). ([26, 15]).

D'autres approches sont possibles par exemple en utilisant des modèles de séparation du signal en deux composantes: une composante Harmonique et une composante "bruit" ([95, 56]). Il est alors possible de modéliser séparément les deux composantes lors de modification de pitch ou de durée. Pour la partie harmonique, il est alors possible de relier les segments en reliant chaque harmonique de part et d'autre de la concaténation. Notons, qu'il est aussi important de modéliser précisément la partie bruit et ainsi de pouvoir conserver une synchronisation temporelle précise entre ces deux composantes. En effet, la composante "bruit" inclut fréquemment une structure temporelle précise dépendante de la fréquence fondamentale (c'est notamment le cas des fricatives voisées ou le bruit de friction généré au niveau d'une constriction du conduit vocal est modulé en intensité par l'ouverture/fermeture périodique de la glotte). Une mauvaise modélisation de cette composante stochastique fait percevoir cette composante comme séparée du reste du signal ([30]).

V.6 Applications

Les applications des systèmes de synthèse à partir du texte ne manquent pas. En voici quelques exemples :

– *Services de télécommunications.* La libéralisation du marché des télécommunications en Europe a récemment rendu les opérateurs de télécommunications plus sensibles au confort de leurs clients. En particulier, on cherche désormais à fournir un maximum de services, à moindre coût. Les synthétiseurs permettent précisément de rendre tout type d'information écrite disponible via le téléphone. On peut ainsi créer des serveurs vocaux diffusant les horaires des cinémas, des informations routières, l'état d'un compte en banque, ou encore des explications automatisées concernant la dernière facture de téléphone. Les requêtes se font soit par la voix (en combinant le synthétiseur avec un reconnaisseur), soit par le clavier du téléphone. AT&T a récemment testé certains services de ce type auprès de ses clients, et constaté un réel engouement, à condition que l'intelligibilité des voix de synthèse

soit suffisante; il s'est avéré que le naturel n'est pas un facteur déterminant pour la plupart de ces services.

- *Apprentissage (ou perfectionnement) de langues étrangères.* Une synthèse de très bonne qualité couplée à un logiciel d'apprentissage constitue un outil très utile à l'apprentissage d'une nouvelle langue, en complément d'un cours avec un professeur. Si ce type de produit n'a pas encore percé sur le marché, c'est à cause de la mauvaise qualité des voix disponibles jusqu'à il y a peu. On voit par contre se multiplier les petits dictionnaires électroniques de poche, qui devraient rapidement être dotés de voix de synthèse. Il en va de même des traducteurs électroniques mot-à-mot qui sont apparus récemment. On pourra par exemple bientôt lire un ouvrage dans un langue étrangère et utiliser un stylo à lecture optique (intégrant un mini-scanner) pour obtenir instantanément la traduction d'un mot inconnu et sa prononciation.
- *Aide aux personnes handicapées.* Les handicaps liés à la parole sont soit d'origine mentale, soit d'origine motrice ou sensorielle. La machine peut être d'un grand secours dans le second cas. Avec l'aide d'un clavier spécialement adapté et/ou d'un logiciel d'assemblage rapide de phrases, un handicapé peut s'exprimer par la voix de son synthétiseur.
- *Livre et jouets parlants.* Le marché du jouet a déjà été touché par la synthèse vocale. De nombreux ordinateurs pour enfants possèdent une sortie vocale qui en augmente l'attrait, particulièrement chez les jeunes enfants (pour qui la voix est le seul moyen de communication avec la machine).
- *Communication homme-machine, multimédia.* A plus long terme, le développement de synthétiseurs de haute qualité (ainsi que la mise au point de reconnaisseurs fiables et robustes) permettra à l'homme de communiquer avec la machine de manière plus naturelle. L'explosion récente du marché du multimédia prouve bien l'intérêt du grand public en la matière.
- *Recherche fondamentale et appliquée.* Enfin, les synthétiseurs possèdent aux yeux des phonéticiens une qualité qui nous fait défaut : ils peuvent répéter deux fois exactement la même chose. Ils sont par conséquent utiles pour la validation de théories relatives à la production, à la perception, ou à la compréhension de la parole.

V.7 Produits

L'offre en produits commerciaux de synthèse à partir du texte s'est considérablement accrue au cours des dernières années. La plupart de ces systèmes sont multilingues, c'est à dire sont capables de produire des voix de synthèse dans plusieurs langues différentes. Ces systèmes incluent, à peu près tous, la synthèse de l'anglais (généralement, américain).

Les configurations matérielles et logicielles diffèrent suivant le type de produits et les applications. La plupart du temps cependant, l'obtention d'une voix de synthèse ne nécessite plus de disposer d'un matériel spécifique (si ce n'est une carte de restitution du son, disponible en standard sur les nouveaux `PC` multimédia), la synthèse proprement dite ne requérant en fait qu'une fraction de la puissance de calcul d'un processeur moderne. Pour certaines applications spécifiques (serveurs vocaux ou applications embarqués), des implantations matérielles sont encore souvent nécessaires.

Parmi les produits les plus importants, citons ceux d'Elan Speech (Elan Sayso, Elan Tempo,...), de Babel Technologies (Brightspeech, Infovox...), d'AT&T (Natural voices,...), de Fonix (Dectalk,...), de Scansoft (Realspeak,...), de Loquendo ou encore IBM.

Une liste de produits plus détaillée peut être trouvée dans [81] ou encore sur le réseau Internet à `http://cslu.cse.ogi.edu/tts/`.

V.8 Conclusion

En 1987, Klatt a écrit " *an articulatory model is likely to be the ultimate solution to the objective of natural intelligible speech synthesis by machine, but computational costs and lack of data upon which to base rules prevent immediate application of this approach.* [54]". Encore aujourd'hui, il n'existe pas réellement de systèmes commerciaux de synthèse vocale à partir de modèles articulatoires en raison d'une part de la complexité des modèles développés et d'autre part de la qualité de synthèse qui reste souvent peu naturelle. L'approche articulatoire recèle pourtant de nombreux avantages. De par son lien direct avec la physiologie, elle apporte des paramètres de contrôle très flexibles et intuitifs pour les modifications de voix (variation de la fréquence fondamentale, taille du conduit vocal, effort d'articulation, etc). Par ailleurs, le potentiel pour la compression très bas débit de la parole est considérable sachant que très peu de paramètres sont alors nécessaires pour synthétiser une voix. D'une façon générale, si on pouvait disposer de méthodes robustes pour l'inversion acoustico-articulatoire (i.e. l'estimation des paramètres articulatoires à partir d'un signal de parole), la modélisation articulatoire pourrait alors constituer la base d'un cadre général incluant la synthèse, le codage et la reconnaissance de parole. Il apparaît donc primordial de poursuivre des recherches dans cette direction. En parallèle, il est aussi essentiel d'améliorer la qualité des synthétiseurs articulatoires.

Malgré le très grand potentiel de l'approche articulatoire, la plupart des applications commerciales dans le domaine de la synthèse reposent sur une modélisation approximative du phénomène de production et suivent une approche traitement de l'information ou traitement du signal. La principale raison est bien entendu liée à la qualité (et la relative simplicité de mise en oeuvre) de cette approche.

Il reste néanmoins plusieurs aspects qui mériteraient d'être améliorés:

- Développer des approches clairement multilingues pour l'analyse du texte (comme le font actuellement les Bell Labs [2]) plutôt que de juxtaposer n systèmes monolingues.
- L'amélioration de la qualité segmentale en utilisant des modèles de décomposition du signal en une composante périodique et une composante apériodique. Il serait en particulier intéressant de parvenir à des méthodes de décomposition plus simples et plus robustes au niveau du bruit.
- Permettre d'effectuer de la synthèse à partir de concepts (c'est à dire qu'ici les phrases exactes sont générées par l'ordinateur et non lues à partir d'un texte.)
- Permettre de changer facilement de voix. Actuellement, si l'on désire avoir plusieurs voix de synthèse, la solution retenue est toujours de ré-enregistrer une nouvelle base de segments de parole. Il serait ainsi extrêmement intéressant de pouvoir modifier la voix de synthèse originale par des techniques de traitement de signal sans avoir à ré-enregistrer une nouvelle base. Il apparaît donc urgent de poursuivre les travaux dans le domaine de la conversion de voix ([50])
- Permettre d'introduire de la variabilité dans la synthèse de parole. En effet, naturellement nous produisons une voix empreinte d'émotion (colère, tristesse, joie,...), et cette émotion joue évidemment un rôle très important dans le naturel de la voix. Il est donc intéressant de pouvoir doter le synthétiseur de telles possibilités.

Bibliographie

[1] Communication de la commission au conseil, au parlement européen au comité économique et social et au comité des régions; vers un espace européen de la recherche. COM(2000) 6, 18 Janvier 2000 (accessible à http://www.cordis.lu), 2000.

[2] R Sproat & al. *Multilingual Text-To-Speech Synthesis - The Bell Labs Approach*. Kluwer Academuic Pub., r. sproat et al. eds. edition.

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

[4] European Linguistic Ressources Association. WWW. http://www.icp.inpg.fr/ELRA/.

[5] Lindblom B. Phonetic contents in phonology. Technical report, Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm (PERILUS) XI, 1990.

[6] Badin. Têtes parlantes audiovisuelles et applications. *http* : *//www.icp.inpg.fr/ badin/TP_Applications.html*, 2001.

[7] L. Bahl and F. Jelinek. Decoding for channels with insertions, deletions adn substitutions with applications to speech recognition. *IEEE on Trans. on Inf. Theory*, IT-21:404–411, 1975.

[8] J. Baker. The dragon system - an overview. *IEEE Trans on ASSP*, 23(1):24–29, Février 1975.

[9] F. Beaugendre. *Une étude perceptive de l'intonation du français*. PhD thesis, Thèse de doctorat de l'Université Paris XI, Orsay, 1994.

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

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

[12] C. Benoit.

[13] A. Black and P. Taylor. Festival speech synthesis system: system documentation. *http://www.cstr.ed.ac.uk/projects/festival.html*, 1997.

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

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

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

[17] G. Burdea and P. Coiffet. *La réalité virtuelle*. Editions hermés edition, 1993. (version en anglais publiée en 1994 sous le titre Virtual reality technology par Wiley Interscience).

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

[19] W. Campbell. *In "TALKING MACHINES: Theories, Models, and Designs."*, chapter Syllable-based Segmental Duration". Elsevier, North Holland, bailly g., benoît c. (editeurs) edition.

[20] Linguistic Data Consortium. WWW. http://www.ldc.upenn.edu.

[21] R. Cox, C. Kamm, L. Rabiner, J. Schoeter, and J. Wilpon. Speech and language processing for next millennium communications services. *Proc IEEE*, 88(8), August 2000.

[22] 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.

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

[24] Van den Heuvel H., Bonafonte A., Boudy J., Dufour S., Lockwood P., Moreno A., and Richard G. Speechdat-car : Towards a collection of speech databases for automotive environment. *in Proc. of Cost 249 Workshop on Speech Recognition Robustness, Finland,*, June 1999.

[25] Van den Heuvel H., Boves L., Moreno A., Omologo M., Richard G., and Sanders E. Annotation in the speechdat projects. *International Journal of Speech Technology*, 2001.

[26] T. Dutoit. *"High quality Text-To-Speech Synthesis of the French Language"*. PhD thesis, Faculté Polytechnique de Mons, 1993.

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

[28] O. Engwall. A 3d tongue model based on mri data. *In Proc of ICSLP 2000.*

[29] O. Engwall. The 3d vocal tract project. http://www.speech.kth.se/multimodal/vocaltract.html.

[30] G. Richard et C. d'Alessandro. "analysis-synthesis of the aperiodic component of speech". *Speech Communication*, 9, Septembre 1996.

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

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

[33] J. Flanagan. Technologies for multimedia communications. *Proc. IEEE*, 84:590–603, 1994.

[34] J. Flanagan, K. Ishizaka, and K. Shipley. Synthesis of speech from a dynamic model of the vocal cords and the vocal tract. *Bell S. T. Journ.*, 54:485–505, 1975.

[35] J. Flanagan, K. Ishizaka, and K. Shipley. Signal models for low bit-rate coding of speech. *J. Acoust. Soc. Am.*, 68(3):780–791, 1980.

[36] Fujisaki. *Proc of ICASSP*, 986.

[37] H. Fujisaki and H. Keikichi. Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of the Acoustical Society of Japan*, 5(4):233–241, 1984.

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

[39] Richard G. The speechdat-car project: Overview of a very large multilingual speech database recorded in cars. *In Proc. of XLDB 2000 (satellite workshop to LREC2000),Athens, Greece,*, May 2000.

[40] Richard G., Goirand M., Sinder D., and Flanagan J.L. Simulation and visualization of articulatory trajectories estimated from speech signals,. *in Proceedings of the International Symposium on Simulation, Visualization and Auralization for Acoustic Research and Education (ASVA97, Tokyo*, April 1997.

[41] M. Gales. *Model-based Techniques for Noise Robust Speech Recognition.* Ph.d. thesis, Cambridge University, Cambridge, UK, 1996.

[42] J-L Gauvain and C-H Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE trans. on Speech and Audio Processing*, 2:291–298, 1994.

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

[44] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *JASA*, 87(4):1738–1752, 1990.

[45] H. Hermansky and N. Morgan. Rasta extensions: robustness to additive and convolutional noise. *Proc. ESCA Workshop on Speech Processing in Adverse conditions*, pages 115–118, 1992.

[46] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE trans. on Speech and Audio Processing*, 2:578–589, 1994.

[47] H. Holma and A. Toskala. *WCDMA for UMTS: Radio access for Third Generation Mobile Communications.* Johne Wiley & Sons Editions, edited by h. holma and a. toskala edition, June 2000.

[48] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara. Revising perceptual linear prediction (plp). *Proc. of Interspeech'05*, pages 2997–3000, 2005.

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

[50] Special issue on Voice Conversion. *Speech Communication*, 16, 1995.

[51] Crothers J. *Typology and universals of vowel systems*, volume Vol. 2: Phonology, chapter in Universals of human language., pages 93–152. edited by J. H. Greenberg, C. A. Ferguson and E. A. Moravcsik (Stanford University Press, Stanford), 1978.

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

[53] D. Klatt. Software for a cascade/parallel formant synthesizer". *J. Acoust. Soc. Amer.*, 67:971–995, 1980.

[54] D. Klatt. "review of text-to-speech conversion for english". *J. Acoust. Soc. Amer.*, 82:737–793, 1987.

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

[56] J. Laroche, E. Moulines, and Y. Stylianou. HNS: Speech modification based on a harmonic + noise model. *Proc. IEEE ICASSP-93, Minneapolis*, Apr 1993.

[57] 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.

[58] M. Lighthill. On sound generated aerodynamically. *Proc. Roy. Soc. London Seires A*, pages 211:564–587, 1952.

[59] J. Liljencrants and B. Lindblom. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, (48):839–862.

[60] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11:215–228, 1992.

[61] Sala M., Sánchez F., Wengelnik H., Van den Heuvel H., Moreno A., Deregibus E., Richard G., and Le Chevalier E. "speechdat-car : Speech databases for voice driven teleservices and control of in-car applications". *in Proc. of EAEC Congress, Spain*, July 1999.

[62] S. Maeda. An aritculatory model based on statistical analysis. *J. Acoust. Soc. Amer.*, S1,S22(A).

[63] F. Malfrere, T. Dutoit, and P. Mertens. Un générateur de parole "tout automatique". *Proc. XXIIèmes JEP*, pages 147–150, 1998.

[64] J. Mariani. *Analyse, synthèse et codage de la parole*. Editions Hermès, 2002.

[65] R. McGowan. Recovering articulator movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary tests. *Speech Com.*, (14):19–48, 1994.

[66] P. Mertens. *in Le français Parlé*, chapter Intonation. Blanche-benveniste,, Paris, editions du cnrs edition.

[67] W. Meyer-Eppler. zum erzeugungsmecahnismus des geräuchlaute. *Z. Phonetik,*, 7:196–212, 1953.

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

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

[70] N.Moreau. *Techniques de compression des signaux*. Collection Technique et Scientifique des télécommunications CNET-ENST. Ed. masson, paris edition, 1995.

[71] O.Engwall. Tongue model. WWW, 2001. http://www.speech.kth.se/ olov/tract.html.

[72] Lieberman P., Crelin E. S., and Klatt D. H. Phonetic ability and related anatomy of the new born, adult human, neanderthal man, and the chimpanzee. *American Anthropologist*, (74):287–307, 1972.

[73] Parthasarathy.

[74] J. Pierrehumbert. "synthesizing intonation". *J. Acoust. Soc. Am*, 70:pp. 985–995, 1981.

[75] 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.

[76] SpeechDat-Car project. WWW. http://www.speechdat.org/SP-CAR/.

[77] SPEECON project. WWW. http://www.speecon.com.

[78] SpeechDat projects. WWW. http://www.speechdat.org.

[79] Carré R., Lindblom B., and MacNeilage P. Rôle de lacoustique dans lévolution du conduit vocal humain. Paris t. 30 série IIb, 471-476., Comptes Rendus de l'Académie des Sciences, 1995.

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

[81] G. Richard and O. Cappé. Synthèse de la parole à partir du texte. *Techniques de l'Ingénieur*, 2003.

[82] Chennoukh S., Sinder D., Richard G., and Flanagan J.L. Improved techniques for voice mimic systems using articulatory codebooks. *in Proc. of EUROSPEECH97*, pages 429–432, 1997.

[83] J. Schroeter and M. Sondhi. Techniques for estimating vocal tract shapes from the speech signal. *IEEE Trans. on Speech and Audio Processing*, 2(1):133–150, Jan. 1994.

[84] R. Sharma, V. Pavlovic, and T. Huang. Toward multimodal human-computer interface. *Proc IEEE*, 86(5), May 1998.

[85] D. Sinder. *Speech synthesis using an aeroacoustic fricative model*. Ph.d. thesis, Rutgers Univ., NJ - USA, 1999.

[86] R. Singh, B. Raj, and R. Stern. *Noise Reduction in Speech applications*, chapter Model Compensation and Matched Condition Models. CRC Press, 2002.

[87] R. Singh, R. Stern, and B. Raj. *Noise Reduction in Speech applications*, chapter Signal and Feature Compensation Methods for Robust Speech Recognition. CRC Press, 2002.

[88] M. Sondhi. Model for wave propagation in a lossy vocal tract. *J. Acoust. Soc. Amer.*, 55(5):1070–1075, 1974.

[89] M. Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Trans. on acoustics Speech and Signal Processing*, ASSP-35(7):955–967, July 1987.

[90] R. Sproat, J Hirshberg, and D. Yarowsky. A corpus based synthesizer. *Proc. of ICSLP*, pages 563–566, 1992.

[91] ETSI Aurora standard. Speech processing, transmission and quality aspects, distributed speech recognition, extended advanced front end feature extraction algorithms,... Technical report, ETSI ES 202 212, v1.1.2, 2005.

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

[93] T.guiard-Marigny, A. Adjoudani, and C. Benoît. *3D models of the lips and Jaw for Visual Speech Synthesis*, chapter 19, pages 247–258. Progress in Speech synthesis. Springer, j.p van santen, r. sproat, j. olive, j. hirschberg edition, 1996.

[94] Traber. *In "TALKING MACHINES: Theories, Models, and Designs."*, chapter "Fo generation with a database of natural Fo patterns and with a neural network". Elsevier, North Holland, bailly g., benoît c. (editeurs) edition.

[95] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on ASSP*, 6(1):1–12, 1998.

Chapitre VI

Notes de cours sur les modèles de Markov Cachés (O. Cappé)

Modèles de mélange et modèles de Markov cachés pour le traitement automatique de la parole

Olivier Cappé

Juin 2000

Contents

1 Modèle de mélange de gaussiennes

1.1 Hypothèses de départ

Dans tout ce document, on considère que la paramétrisation du signal de parole a été effectuée au préalable. C'est à dire que les données à modéliser se présentent directement sous la forme vecteurs de paramètres (il s'agit en général de coefficients cepstraux) observés à intervalle régulier. Pour fixer les ordres de grandeur, chaque vecteur de paramètres comporte typiquement de 10 à 30 paramètres mesurés toutes les 10 ms.

Le **modèle de mélange** [12] consiste à supposer que les vecteurs observés $\mathbf{X}_t$ sont des réalisations de variables aléatoires mutuellement indépendantes, qui suivent toute une même loi ayant la forme suivante [1]

$$f(\mathbf{x}) = \sum_{i=1}^{M} \pi_i f_i(\mathbf{x}) \tag{1}$$

où chaque $f_i(\mathbf{x})$ est une densité de probabilité, et les π_i sont des scalaires positifs. Le fait que $f()$ soit une densité de probabilité implique que $\sum_{i=1}^{M} \pi_i = 1$. On peut très bien considérer que (1) définit une stratégie de type régression dans laquelle on cherche à décomposer une fonction inconnue et a priori complexe $f()$ sur un ensemble de fonctions plus simples $f_i()$. Nous allons voir cependant que ce modèle peut être interprété d'une manière plus pertinente en supposant que les données observées sont réparties dans différentes **classes** (on parle aussi de **composantes du mélange**).

1.2 Composantes du mélange

Imaginons que les vecteurs d'observations $\mathbf{X}$ soient générés par le mécanisme suivant : *(i)* on tire une variable aléatoire discrète s à valeur dans $1, \ldots, M$ où M désigne le nombre de classes du mélange, on note $\pi_i = P\{s = i\}$ pour $i = 1, \ldots, M$ les probabilités respectives de tirer chacune des classes ; *(ii)* Conditionnellement à $\{S = i\}$, $\mathbf{X}$ est distribué selon la loi de densité de probabilité $f_i(\mathbf{x})$. La densité conditionnelle du vecteur $\mathbf{X}$ peut s'écrire sous la forme concise suivante

$$g(\mathbf{x}|s) = \sum_{i=1}^{M} f_i(\mathbf{x}) \mathbb{I}_{(S=i)} \tag{2}$$

où $\mathbb{I}$ désigne la fonction indicatrice d'un événement. La densité conjointe du vecteur $\mathbf{X}$ et de l'indicatrice S de la composante du mélange s'écrit en vertu de la règle de Bayes

$$h(\mathbf{x}, s) = g(\mathbf{x}|s)p(s) = \left(\sum_{i=1}^{M} f_i(\mathbf{x}) \mathbb{I}_{(S=i)} \right) \left(\sum_{i=1}^{M} \pi_j \mathbb{I}_{(S=j)} \right) = \sum_{i=1}^{M} \pi_i f_i(\mathbf{x}) \mathbb{I}_{(S=i)} \tag{3}$$

La loi marginale du vecteur observé $\mathbf{X}$ s'obtient en sommant (3) sur le domaine de S

$$f(\mathbf{x}) = \sum_{i=1}^{M} \pi_i f_i(\mathbf{x}) \tag{4}$$

[1] Dans ce document, les vecteurs $\mathbf{a}$ sont par convention toujours des vecteurs colonne (par exemple, la norme L^2 du vecteur s'écrit $\sqrt{\mathbf{a}'\mathbf{a}}$).

Ce qui correspond bien sûr à l'équation (4) qui définit le modèle de mélange. Le modèle de mélange est donc équivalent à un modèle dans lequel on suppose que les données sont réparties aléatoirement (et indépendamment les unes des autres) en M classes qui sont chacune caractérisée par une distribution différente $f_i()$.

Comme le laisse présager ce qui précède, la variable indicatrice S est une donnée constitutive du problème qui présente l'inconvénient de ne pouvoir être observée en pratique : on observe des réalisations du vecteur aléatoire $\mathbf{X}$ sans savoir de manière certaine quelle est la classe du mélange associée à chaque observation. Au sens de l'algorithme EM, la variable S constitue une **donnée latente**, c'est-à-dire fortement suggérée par le problème considéré (on parle également de donnée non-observée ou manquante). Nous verrons que l'introduction de ces données non-observées permet de résoudre de manière élégante un problème d'estimation relativement complexe.

L'utilisation d'un tel modèle pour les vecteurs de paramètres issus de l'analyse de signaux de parole se justifie essentiellement en faisant appel à l'interprétation des classes du mélange développée ci-dessus : pour la parole, il n'est pas stupide de supposer que les vecteurs de paramètres (qui d'une certaine façon représentent le spectre à court-terme du signal) vont se répartir différemment selon les caractéristique du son de parole considéré (son voisé / non voisé, ou plus finement en fonction du phonème). L'hypothèse qui peut paraître la plus réductrice ici est le fait qu'on suppose que les vecteurs observés sont indépendants. Cette hypothèse est acceptable dans des applications telles que la reconnaissance du locuteur en mode "texte libre" dans la mesure où le texte prononcé par le locuteur et totalement inconnu et que l'on n'exerce aucun contrôle sur cet aspect des choses. Au contraire, dans des applications où le texte prononcé est fixé (pour la reconnaissance de la parole par exemple), la prise en compte de l'aspect séquentiel (c'est-à-dire de l'ordre dans lequel les vecteurs sont observés) constitue un élément primordial. Nous verrons plus précisément au paragraphe 4 dans quelle mesure les modèles de Markov cachés permettent de modéliser cet aspect des données.

1.3 Rappel concernant les densités gaussiennes multidimensionnelles

Dans la suite, on supposera que les densités $f_i(\mathbf{x})$ correspondent à des loi normales multidimensionnelles et on notera $\boldsymbol{\mu}_i$ et $\boldsymbol{\Sigma}_i$ les vecteurs moyens et les matrices de covariance correspondants (pour $1 \leq i \leq M$). Les matrices de covariance seront supposées inversibles. On rappelle que $f_i(\mathbf{x})$ est alors décrite par l'expression analytique suivante

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \left|\boldsymbol{\Sigma}_i\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right] \tag{5}$$

où p désigne la dimension des vecteurs.
Exercice : Vérifiez que l'on retrouve bien la forme connue dans le cas $p = 1$. Vérifiez de plus que l'écriture dans le cas $p = 1$ est cohérente avec l'écriture ci-dessus dans le cas où l'on suppose que les coordonnées du vecteur $\mathbf{x}$ sont mutuellement indépendantes. Comment se traduit l'hypothèse d'indépendance sur la structure de la matrice de covariance $\boldsymbol{\Sigma}_i$ (on pourra se reporter à [3] si l'exercice résiste vraiment).

Il est un peu difficile de se figurer l'allure d'une gaussienne en dimension $p > 1$, toutefois, on remarque que si l'on suppose l'existence d'une matrice $\mathbf{R}_i$ telle que $\boldsymbol{\Sigma}_i^{-1} =$

$\mathbf{R}_i'\mathbf{R}_i$, on note que $f_i(\mathbf{x}) = C^{te}$ est équivalent à

$$\mathbf{y}'\mathbf{y} = C^{te} \quad \text{avec} \quad \mathbf{y} = \mathbf{R}_i(\mathbf{x} - \boldsymbol{\mu}_i) \tag{6}$$

Cette équation définit une sphère dans $\mathbb{R}^p$ (on parle plutôt d'hypersphère lorsque $p > 3$). Sachant que $\mathbf{x} = \boldsymbol{\mu}_i + \mathbf{R}_i^{-1}\mathbf{y}$, la surface $f_i(\mathbf{x}) = C^{te}$ s'obtient à partir de cette hypersphère par un changement de base suivit d'un décalage par le vecteur $\boldsymbol{\mu}_i$. On obtient ce que l'on appelle un ellipsoïde (qui se réduit à une ellipse en dimension $p = 2$), centré autour du vecteur moyen $\boldsymbol{\mu}_i$, et dont la forme exacte dépend des caractéristiques de la matrice de covariance $\boldsymbol{\Sigma}_i$.
Remarque : La matrice $\mathbf{R}_i$ existe toujours, il s'agit de la décomposition dite de Cholesky [5] d'une matrice positive.

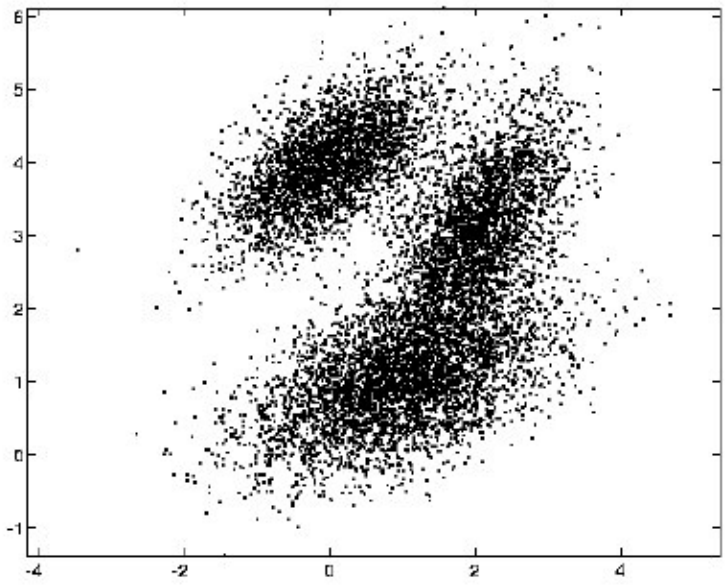


Figure 1: Mille points issus d'un mélange de trois densités gaussiennes bidimensionnelles

A titre d'illustration, la figure 1.3 présente un exemple en deux dimensions dans lequel on a représenté mille points issus d'un modèle de mélange comportant trois composantes. On distingue distinctement les trois forme ellipsoïdales (qui n'ont d'ailleurs pas la même orientation, ce qui traduit le fait que les matrices de covariances sont différentes pour chaque composante du mélange). Il est clair que selon la manière dont les vecteurs moyens des composantes diffèrent, l'individualisation des composantes sera plus où moins marquée. Le poids de chaque composante du mélange π_i se traduit par la proportion statistique de vecteurs venant de chacune des composantes (par application de la loi des grands nombres). Autant que l'on puisse en juger d'après ce tracé, les poids sont ici à peu près identiques pour les trois composantes (c'est-à-dire voisins de 1/3).

2 L'algorithme EM

2.1 Difficultés posées par l'estimateur du maximum de vraisemblance

On ne considère ici que l'estimation des paramètres du modèle de mélange au sens du maximum de vraisemblance. On sait en effet que cette stratégie d'estimation conduit à

des estimateurs asymptotiquement efficace, c'est-à-dire "optimaux" lorsque le nombre de données observées devient important [2]. Le problème est que la fonction de vraisemblance prend ici la forme relativement complexe suivante

$$F(\mathcal{X};\boldsymbol{\Theta}) = \prod_{t=1}^{T}\sum_{i=1}^{M}\pi_i f_i(\mathbf{x}_t) \tag{7}$$

où $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_T\}$ désigne l'ensemble des T vecteurs dont on dispose pour estimer les paramètres du modèle, et $\boldsymbol{\Theta}$ regroupe les paramètres du modèle à estimer, c'est-à-dire π_i, $\boldsymbol{\mu}_i$ et $\boldsymbol{\Sigma}_i$ pour $i = 1, \ldots, M$. Dans cette équation, les densités $f_i()$ doivent être remplacées par leur expression donnée à l'équation (5).

On conçoit aisément qu'il y ait quelques difficulté à maximiser la fonction de vraisemblance définie par (7) par rapport aux paramètres $\boldsymbol{\Theta}$ du modèle. Outre une expression analytique plutôt complexe, cette fonction de vraisemblance présente le défaut de ne pas être convexe, c'est-à-dire qu'elle n'admet a priori pas de maximum unique [9]. Par conséquent même l'utilisation d'algorithmes d'optimisation classiques (gradient, Newton) est ici délicate car il faudrait auparavant cerner le domaine dans lequel on désire rechercher la valeur optimale de $\boldsymbol{\Theta}$, chose qui est plutôt malaisée lorsque $\boldsymbol{\Theta}$ est un paramètre de dimension très élevée comportant des données non-homogènes (par exemple les poids π_i et les matrices de covariance $\boldsymbol{\Sigma}_i$) [9].

2.2 Principe de l'algorithme EM

L'algorithme EM apporte une solution extrêmement générale à ce type de problèmes pour lesquels le modèle statistique considéré peut être complété en faisant appel à des données latentes. Dans le cas du modèle de mélange nous avons vu que la variable non observée qu'il est judicieux de faire intervenir est l'indicatrice de la composante du mélange S_t associée à chaque vecteur observé $\mathbf{x}_t$. Si ces données latentes pouvaient être observées, la solution du problème serait beaucoup plus simple.

L'idée à la base de l'algorithme EM consiste à raisonner sur les données complètes (données observées et données latentes) tout en prenant en compte le fait que l'information disponible sur les données latentes ne peut venir que des données observées. Ceci se traduit par un algorithme itératif pour lequel chaque itération se décompose deux étapes [4]

Expectation

$$\boxed{Q_{\boldsymbol{\Theta}_n}(\boldsymbol{\Theta}) = E\left[\log H(\mathcal{X},\mathcal{S};\boldsymbol{\Theta})\middle|\,\mathcal{X};\boldsymbol{\Theta}_n\right]} \tag{8}$$

Maximization

$$\boxed{\boldsymbol{\Theta}_{n+1} = \arg\max_{\boldsymbol{\Theta}} Q_{\boldsymbol{\Theta}_n}(\boldsymbol{\Theta})} \tag{9}$$

où la notation $\boldsymbol{\Theta}_n$ désigne la valeur estimée des paramètres du modèle à la $n$ième itération de l'algorithme, $\mathcal{X}$ et $\mathcal{S}$ désignent respectivement l'ensemble des données observées et des données latentes associées, enfin, $H()$ désigne la vraisemblance conjointe des données observées et latentes.

2.3 Propriétés de l'algorithme EM

La première et la plus importante propriété de l'algorithme EM et le fait que la suite des valeurs estimées $\{\mathbf{\Theta}_n\}$ est construite de façon à ce que la vraisemblance des données observées augmente à chaque itération de l'algorithme. En effet

$$Q_{\mathbf{\Theta}_n}(\mathbf{\Theta}_{n+1}) - Q_{\mathbf{\Theta}_n}(\mathbf{\Theta}_n) = E\left[\log H(\mathcal{X},\mathcal{S};\mathbf{\Theta}_{n+1})\middle|\,\mathcal{X};\mathbf{\Theta}_n\right] - E\left[\log H(\mathcal{X},\mathcal{S};\mathbf{\Theta}_n)\middle|\,\mathcal{X};\mathbf{\Theta}_n\right] \quad (10)$$

Ce qui se réécrit en faisant intervenir la vraisemblance conditionnelle $K()$ (données latentes sachant les données observées) et la vraisemblance des données observées $F()$, et en utilisant la règle de Bayes

$$Q_{\mathbf{\Theta}_n}(\mathbf{\Theta}_{n+1}) - Q_{\mathbf{\Theta}_n}(\mathbf{\Theta}_n) = \log F(\mathcal{X};\mathbf{\Theta}_{n+1}) - \log F(\mathcal{X};\mathbf{\Theta}_n) + E\left[\log \frac{K(\mathcal{S}|\mathcal{X};\mathbf{\Theta}_{n+1})}{K(\mathcal{S}|\mathcal{X};\mathbf{\Theta}_n)}\middle|\,\mathcal{X};\mathbf{\Theta}_n\right] \quad (11)$$

Le second terme du membre de droite de l'équation s'écrit sous sa forme intégrale

$$\int \log \frac{K(\mathcal{S}|\mathcal{X};\mathbf{\Theta}_{n+1})}{K(\mathcal{S}|\mathcal{X};\mathbf{\Theta}_n)} K(\mathcal{S}|\mathcal{X};\mathbf{\Theta}_n)\mathbf{d}\mathcal{S} \quad (12)$$

Cette dernière quantité est du type $\int \log[k_2(z)/k_1(z)]k_1(z)dz$, c'est-à-dire qu'elle est toujours négative, en vertu de l'inégalité de Jensen qui indique que pour toute fonction convexe $t()$, $t(E[k(Z)]) \leq E[t \circ k(Z)]$ (Sous réserve d'intégrabilité des quantités figurant à l'intérieur du signe $E[]$). En appliquant cette inégalité avec la fonction $t() = -\log()$, on vérifie facilement l'affirmation précédente (Remarque : ce type de quantité joue un rôle important en estimation statistique en tant que mesure de la similarité entre deux densités de probabilité, il s'agit de la divergence de Kullback). Par conséquent, dans la mesure où le membre de gauche de l'équation (11) est positif (car $\mathbf{\Theta}_{n+1}$ maximise $Q_{\mathbf{\Theta}_n}(\mathbf{\Theta})$), la première quantité figurant au second membre de (11) est nécessairement positive, c'est-à-dire que

$$\boxed{\log F(\mathcal{X};\mathbf{\Theta}_{n+1}) \geq \log F(\mathcal{X};\mathbf{\Theta}_n)} \quad (13)$$

La première vertu de l'algorithme EM est donc de permettre de construire une suite d'estimateurs des paramètres du modèle pour laquelle la vraisemblance croît. Bien sûr, ce qui rend l'algorithme EM intéressant en pratique est le fait que les deux équations (8) et (9) vont avoir une forme analytique explicite, et ce pour une classe très large de modèles statistiques. Ce point est détaillé pour les modèles de mélange, puis des modèles de Markov cachés conditionnellement gaussiens aux paragraphes 3 et 6.

La seconde propriété importante de l'algorithme EM et que si l'on différencie la quantité intermédiaire par rapport au paramètre $\mathbf{\Theta}$, il vient

$$\left.\frac{\partial Q_{\mathbf{\Theta}_n}(\mathbf{\Theta})}{\partial \mathbf{\Theta}}\right|_{\mathbf{\Theta}=\mathbf{\Theta}_n} = \left.\frac{\partial \log F(\mathcal{X};\mathbf{\Theta})}{\partial \mathbf{\Theta}}\right|_{\mathbf{\Theta}=\mathbf{\Theta}_n} + \left.\frac{\partial E\left[\log K(\mathcal{S}|\mathcal{X};\mathbf{\Theta})\middle|\,\mathcal{X};\mathbf{\Theta}_n\right]}{\partial \mathbf{\Theta}}\right|_{\mathbf{\Theta}=\mathbf{\Theta}_n} \quad (14)$$

Le dernier terme au second membre de l'équation s'écrit dans sous une forme intégrale du type

$$\left.\frac{\partial \int \log k(z;\phi) k(z;\phi_0) dz}{\partial \phi}\right|_{\phi_0} = \left.\frac{\partial \int k(z;\phi) dz}{\partial \phi}\right|_{\phi_0}$$

ce en supposant que la fonction $k()$ est suffisamment régulière pour permettre la permutation de la dérivation par rapport à ϕ et l'intégration par rapport à z. Ce terme est donc nul puisque $\int k(z;\phi)dz = 1$ pour toutes les valeurs du paramètre ϕ si $k()$ est une famille de densités de probabilité. Par conséquent (14) implique que

$$\boxed{\left.\frac{\partial Q_{\boldsymbol{\Theta}_n}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_n} = \left.\frac{\partial \log F(\mathcal{X};\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_n}} \tag{15}$$

L'algorithme EM possède donc une seconde vertu, en ce qu'il permet de calculer le gradient de la fonction d'objectif (la vraisemblance) aux points $\boldsymbol{\Theta}_n$. Cette propriété est très importante pour la convergence de l'algorithme puisque qu'elle implique que les points stables de l'algorithme (c'est-à-dire des points tels que $Q_{\boldsymbol{\Theta}_*}(\boldsymbol{\Theta})$ soit maximum en $\boldsymbol{\Theta}_*$) sont des points stationnaires de la vraisemblance pour lesquels $\left.\frac{\partial \log F(\mathcal{X};\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}_*} = \mathbf{0}$.

Avec quelques hypothèses concernant la régularité de la fonction intermédiaire $Q()$ et la structure topologique de l'espace des paramètres $\boldsymbol{\Theta}$, (15) permet de montrer que l'algorithme EM ne peut converger que vers des points stationnaires de la vraisemblance. En pratique, la vraisemblance n'étant pas une fonction convexe des paramètres $\boldsymbol{\Theta}$, c'est-à-dire pouvant présenter plusieurs maximums locaux, le comportement de l'algorithme EM dépend fortement de la valeur initiale $\boldsymbol{\Theta}_1$ depuis laquelle on le fait démarrer.

3 Algorithme EM pour les mélanges de gaussiennes

3.1 Quantité intermédiaire de l'algorithme EM

Pour le modèle de mélange, nous avons déjà rencontré la densité conjointe des données observées et des données latentes à l'équation (3). On en déduit que la log-vraisemblance des données complètes s'écrit

$$\log F(\mathcal{X}, \mathcal{S}; \boldsymbol{\Theta}) = \sum_{t=1}^{T} \sum_{i=1}^{M} \log(\pi_i f_i(\mathbf{x}_t)) \mathbb{I}_{(S_t=i)} \tag{16}$$

Dans cette équation le logarithme est passé à l'intérieur de la sommation sur l'indice i uniquement car cette somme se réduit à un seul terme du fait de la présence des fonctions indicatrices. La quantité intermédiaire de l'algorithme EM s'écrit donc

$$Q_{\boldsymbol{\Theta}_n}(\boldsymbol{\Theta}) = \sum_{t=1}^{T} \sum_{i=1}^{M} \log(\pi_i f_i(\mathbf{x}_t)) E[\mathbb{I}_{(S_t=i)} | \mathcal{X}; \boldsymbol{\Theta}_n] \tag{17}$$

Le dernier terme de cette équation est par définition égal à $P(S_t = i | \mathbf{X}, \boldsymbol{\Theta}_n)$. Dans la mesure où l'on a supposé les observations indépendantes, ce terme peut se réécrire

plus simplement sous la forme $P(S_t = i|\mathbf{x}_t; \mathbf{\Theta}_n)$. Le calcul de la quantité intermédiaire de l'algorithme EM se réduit donc au calcul de $P(S_t = i|\mathbf{x}_t; \mathbf{\Theta}_n)$ pour $1 \leq i \leq M$ et $1 \leq t \leq T$. Ces quantités, que l'on notera $\gamma_t^{(n)}(i)$ pour simplifier les écritures, peuvent être calculées simplement par application de la formule de Bayes

$$\gamma_t^{(n)}(i) = P(S_t = i|\mathbf{x}_t; \mathbf{\Theta}_n) = \frac{g(\mathbf{x}_t|S_t = i)P\{S_t = i\}}{f(\mathbf{x}_t)} = \frac{g(\mathbf{x}_t|S_t = i)P\{S_t = i\}}{\sum_{j=1}^{M} g(\mathbf{x}_t|S_t = j)P\{S_t = j\}}$$
$$= \frac{\pi_i f_i(\mathbf{x}_t)}{\sum_{j=1}^{M} \pi_j f_j(\mathbf{x}_t)} \quad (18)$$

Etant entendu que dans toutes les quantités figurées à droite du premier signe égal, les paramètres considérés sont ceux obtenus à l'itération d'indice n. En remplaçant les densités $f_i()$ par leur expression analytique, on obtient

$$\boxed{\gamma_t^{(n)}(i) = \frac{\pi_i \left|\mathbf{\Sigma}_i\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i)\right]}{\sum_{j=1}^{M} \pi_j \left|\mathbf{\Sigma}_j\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_j)'\mathbf{\Sigma}_j^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_j)\right]}} \quad (19)$$

Même remarque que ci-dessus, il faudrait à droite du signe égal rajouter l'exposant (n) au dessus de chacun des paramètres du modèle, ce que l'on omet pour des questions de lisibilité. En insérant ces valeurs dans (17), on obtient

$$Q_{\mathbf{\Theta}_n}(\mathbf{\Theta}) = \sum_{t=1}^{T}\sum_{i=1}^{M} \gamma_t^{(n)}(i) \left[\log \pi_i + \log f_i(\mathbf{x}_t))\right] \quad (20)$$

soit compte tenu de (5)

$$Q_{\mathbf{\Theta}_n}(\mathbf{\Theta}) = -\frac{Tp}{2}\log 2\pi + \sum_{t=1}^{T}\sum_{i=1}^{M} \gamma_t^{(n)}(i) \left\{\log \pi_i - \frac{1}{2}\left[\log\left|\mathbf{\Sigma}_i\right| + (\mathbf{x}_t - \boldsymbol{\mu}_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i)\right]\right\} \quad (21)$$

3.2 Réestimation des paramètres

C'est la maximisation de (21) par rapport aux paramètres π_i, $\boldsymbol{\mu}_i$ et $\mathbf{\Sigma}_i$ qui fournit les nouvelles valeurs estimées pour l'itération $(n+1)$. Le problème est que cette quantité dépend de paramètres vectoriels $(\boldsymbol{\mu}_i)$, voire matriciels $(\mathbf{\Sigma}_i)$.

Le cas des poids des composantes de mélange π_i est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit il faut tenir compte de la contrainte qui existe sur ces paramètres : $\sum_{i=1}^{M} \pi_i = 1$ (car les π_i correspondent à une loi de probabilité discrète). La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de

Lagrange associé à cette contrainte. Et l'on obtient

$$\boxed{\pi_i^{(n+1)} = \frac{\sum_{t=1}^{T} \gamma_t^{(n)}(i)}{\sum_{j=1}^{M}\sum_{t=1}^{T} \gamma_t^{(n)}(j)} = \frac{1}{T}\sum_{t=1}^{T} \gamma_t^{(n)}(i)} \tag{22}$$

En ce qui concerne les autres paramètres, la maximisation est un peu plus délicate. La méthode générique pour aborder ce type de problèmes passe par la définition de la *différentielle* d'une fonction $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ (fonction d'une matrice de dimension (p, q) à valeur dans $\mathbb{R}$). On peut dans certains cas s'en passer dans la mesure où la différentielle, lorsqu'elle existe, s'identifie avec la matrice des dérivées partielles $[\partial f(\mathbf{M})/\partial m_{ij}]$ de la fonction [6].

Cette interprétation "simpliste" (calcul de la différentielle élément par élément) suffit ici pour venir à bout des termes linéaires (pour $\boldsymbol{\mu}_i$ notament), mais il reste le problème des termes qui font intervenir $\log|\boldsymbol{\Sigma}_i|$ et $\boldsymbol{\Sigma}_i^{-1}$. Pour simplifier les calculs, on peut réécrire (21) en fonction de $\boldsymbol{\Phi}_j = \boldsymbol{\Sigma}_j^{-1}$ (matrice dite de précision) [1]. Ce changement de paramétrisation permet d'éviter les termes en $\boldsymbol{\Sigma}_i^{-1}$ et, en ce qui concerne $\log|\boldsymbol{\Phi}_j| = -\log|\boldsymbol{\Sigma}_i|$, il s'agit d'un résultat classique :

Exercice : Vérifiez par le calcul composante par composante que

$$\frac{\partial(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Phi}(\mathbf{x}-\boldsymbol{\mu})}{\partial\boldsymbol{\mu}} = -(\boldsymbol{\Phi}+\boldsymbol{\Phi}')(\mathbf{x}-\boldsymbol{\mu})$$

$$\frac{\partial(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Phi}(\mathbf{x}-\boldsymbol{\mu})}{\partial\boldsymbol{\Phi}} = (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'$$

$$\frac{\partial\log|\boldsymbol{\Phi}|}{\partial\boldsymbol{\Phi}} = \left(\boldsymbol{\Phi}^{-1}\right)'$$

Dans le cas particulier considéré ici, la première et la troisième relation se simplifient en utilisant le fait que $\boldsymbol{\Phi}$ est symétrique. Pour vérifier la dernière relation il faut utiliser le fait que le déterminant d'une matrice et son inverse peuvent s'écrire en fonction des cofacteurs (expansion dite de Laplace du déterminant). En déduire les deux dernières équations de réestimation des paramètres du modèle de mélange données ci-dessous.

Tous calculs faits, on obtient les équations de réestimation suivantes pour les vecteurs moyens et les matrices de covariance des composantes du mélange

$$\boxed{\boldsymbol{\mu}_i^{(n+1)} = \frac{\sum_{t=1}^{T} \gamma_t^{(n)}(i)\mathbf{x}_t}{\sum_{t=1}^{T} \gamma_t^{(n)}(i)}} \tag{23}$$

$$\boxed{\boldsymbol{\Sigma}_i^{(n+1)} = \frac{\sum_{t=1}^{T} \gamma_t^{(n)}(i)(\mathbf{x}_t-\boldsymbol{\mu}_i^{(n+1)})(\mathbf{x}_t-\boldsymbol{\mu}_i^{(n+1)})'}{\sum_{t=1}^{T} \gamma_t^{(n)}(i)}} \tag{24}$$

4 Modèle de Markov caché

Dans toute cette partie, on considère l'extension des résultats précédents au cas des modèles de Markov cachés[2]. Les références bibliographiques à consulter sont en priorité [8], ainsi que [11] ou [10].

4.1 Définition

Le modèle de Markov caché partage avec le modèle de mélange la caractéristique essentielle de faire intervenir une structure sous-jacente (non observable) sous la forme d'une variable indicatrice (ou étiquette), associée à chaque observation, et prenant un nombre fini de valeurs. Le modèle de Markov caché est toutefois plus riche que le modèle de mélange dans le sens où il permet de rendre compte des interactions temporelles en substituant à l'hypothèse d'indicatrices *iid* celle d'une évolution markovienne.

Plus précisément, on dira qu'un processus aléatoire $\{\mathbf{X}_t\}_{t\geq 1}$ (éventuellement vectoriel) a une structure de *modèle de Markov caché*, si il existe un processus aléatoire $\{S_t\}_{t\geq 1}$ (défini sur le même espace de probabilité), prenant un nombre fini N de valeurs, tel que :

1. Les indicatrices S_t on une évolution markovienne "homogène" (c'est-à-dire indépendante de l'indice temporel)

$$p(s_t|s_{1:t-1}) = p(s_t|s_{t-1}) = p(s_2|s_1) \tag{25}$$

   où la notation $s_{1:t-1}$ désigne la séquence $\{s_1, s_2, \ldots, s_{t-1}\}$.

2. Les observations $\mathbf{X}_t$ sont indépendantes conditionnellement aux indicatrices S_t

$$p(\mathbf{x}_{1:T}|s_{1:T}) = \prod_{t=1}^{T} p(\mathbf{x}_t|s_t) \tag{26}$$

La manière usuelle de paramétrer un tel modèle consiste[3]

- pour la partie markovienne, à spécifier la *distribution initiale* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$ où $\pi_i = P(S_1 = i)$ et la matrice de transition $\mathbf{A} = [a_{ij}]$ où $a_{ij} = P(S_{t+1} = j|S_t = i)$. On note que la matrice $\mathbf{A}$ possède une structure particulière, dite de matrice stochastique, pour laquelle $\sum_{j=1}^{N} a_{ij} = 1$ (pour toutes les lignes).
- pour la partie observation, la loi $f_i(\mathbf{x}_t) = p(\mathbf{x}_t|S_t = i)$ appartient en général à une même famille paramétrique de paramètre $\boldsymbol{\theta}_i$. Dans le cas particulier qui va nous retenir, celui des HMM conditionnellement gaussien, la loi $f_i(\mathbf{x})$ est une loi normale multivariée paramétrée par son vecteur moyen $\boldsymbol{\mu}_i$ et sa matrice de covariance $\boldsymbol{\Sigma}_i$.

[2]Dans la suite nous utiliserons plutôt l'appellation anglaise de *HMM* pour *Hidden Markov Model* qui est la plus largement employée.

[3]On notera que dans cette partie du texte, on renonce à différencier explicitement les densités de probabilités : celles-ci sont toutes notées $p()$.

4.2 Propriétés élémentaires

On établit dans cette partie quelques propriétés fort utiles des HMM. Notons tout d'abord que

$$p(\mathbf{x}_t, s_t|\mathbf{x}_{t-1}, s_{t-1}) = \frac{p(\mathbf{x}_{t-1:t}|s_{t-1:t})p(s_{t-1:t})}{p(\mathbf{x}_{t-1}|s_{t-1})p(s_{t-1})} = p(\mathbf{x}_t|s_t)p(s_t|s_{t-1}) \tag{27}$$

et que de même

$$p(\mathbf{x}_t, s_t|\mathbf{x}_{1:t}, s_{1:t}) = \frac{p(\mathbf{x}_{1:t}|s_{1:t})p(s_{1:t})}{p(\mathbf{x}_{1:t-1}|s_{1:t-1})p(s_{1:t-1})} = p(\mathbf{x}_t|s_t)p(s_t|s_{t-1}) \tag{28}$$

c'est à dire que le processus joint $\{\mathbf{X}_t, S_t\}$ est markovien homogène, tout comme $\{S_t\}$, à la seule différence que son espace d'états (l'espace dans lequel il prend ses valeurs) n'est pas fini. Par contre, il est important de garder à l'esprit le fait que le processus observé $\{\mathbf{X}_t\}$ seul n'est pas markovien puisque

$$p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) = \sum_{i=1}^{N} p(\mathbf{x}_t, S_t = i|\mathbf{x}_{1:t-1}) = \sum_{i=1}^{N} p(\mathbf{x}_t|S_t = i)P(S_t = i|\mathbf{x}_{1:t-1}) \tag{29}$$

Cette dernière équation montre bien que la loi de $\mathbf{X}_t$ conditionnellement à son passé est un modèle de mélange dont les poids $P(S_t = i|\mathbf{x}_{1:t-1})$ dépendent du passé complet du signal (et pas seulement de $\mathbf{X}_{t-1}$).

Nous rappelons par ailleurs un résultats classique des processus markoviens qui est que

$$p(f(s_{t_1:t_2}), h(s_{t_4:t_5})|g(s_{t_3})) = p(f(s_{t_1:t_2})|g(s_{t_3}))\, p(h(s_{t_4:t_5})|g(s_{t_3})) \tag{30}$$

dès que $t_1 \leq t_2 \leq t_3 < t_4 \leq t_5$ ou $t_1 \leq t_2 < t_3 \leq t_4 \leq t_5$ (où f, g et h sont des fonctions mesurables). Ce qu'on résume souvent en disant que le passé et le future d'une chaîne de Markov sont conditionnellement indépendant lorsque l'on conditionne par rapport au point courant.

5 Procédures de filtrage

5.1 Filtre d'état

Dans cette partie, nous nous intéressons aux procédures qui permettent de calculer pratiquement (avec un coût de calcul raisonnable) des quantités telles que celles qui apparaissent dans l'équation (29). La première solution intuitive basée sur le conditionnement se révèle impraticable, en effet

$$\begin{aligned} p(s_t|\mathbf{x}_{1:t-1}) &= \frac{\displaystyle\sum_{1\leq s_1,\ldots,s_{t-1}\leq N} p(s_{1:t}, \mathbf{x}_{1:t-1})}{p(\mathbf{x}_{1:t-1})} \\ &= \frac{\displaystyle\sum_{1\leq s_1,\ldots,s_{t-1}\leq N} \prod_{r=1}^{t-1} p(\mathbf{x}_r|s_r) \prod_{r=1}^{t-1} p(s_{r+1}|s_r)p(s_1)}{p(\mathbf{x}_{1:t-1})} \end{aligned} \tag{31}$$

où le dénominateur s'obtient par simple marginalisation (sommation pour toutes les valeurs de s_t) du numérateur. L'évaluation de cette expression, bien que théoriquement possible, présente un coût de calcul prohibitif lorsque la longueur de la séquence observée augmente puisque le nombre de termes impliqués dans la somme figurant au numérateur est N^{t-1}.

En fait il existe une solution de calcul récursive dont la complexité est en $N^2(t-1)$, donc simplement proportionnelle à la longueur de la séquence. Notons $\phi_t(i) \triangleq P(S_t = i|\mathbf{x}_{1:t-1})$, où $\phi_t(1) = \pi_i$ par définition.

$$\begin{aligned}
\phi_t(j) &= \frac{\sum_{i=1}^{N} p(S_t = j, S_{t-1} = i, \mathbf{x}_{1:t-1})}{p(\mathbf{x}_{1:t-1})} = \frac{\sum_{i=1}^{N} p(S_t = j, \mathbf{x}_{1:t-1}|S_{t-1} = i)p(S_{t-1} = i)}{p(\mathbf{x}_{1:t-1})} \\
&= \frac{\sum_{i=1}^{N} P(S_t = j|S_{t-1} = i)p(\mathbf{x}_{1:t-1}|S_{t-1} = i)P(S_{t-1} = i)}{p(\mathbf{x}_{1:t-1})} \\
&= \frac{\sum_{i=1}^{N} a_{ij}p(\mathbf{x}_{t-1}|S_{t-1} = i)p(\mathbf{x}_{1:t-2}, S_{t-1} = i)}{p(\mathbf{x}_{1:t-1})} = \frac{\sum_{i=1}^{N} \phi_{t-1}(i)f_i(\mathbf{x}_{t-1})a_{ij}}{p(\mathbf{x}_{t-1}|\mathbf{x}_{1:t-2})} \\
&= \frac{\sum_{i=1}^{N} \phi_{t-1}(i)f_i(\mathbf{x}_{t-1})a_{ij}}{\sum_{k=1}^{N}\sum_{i=1}^{N} \phi_{t-1}(i)f_i(\mathbf{x}_{t-1})a_{ik}} = \frac{\sum_{i=1}^{N} \phi_{t-1}(i)f_i(\mathbf{x}_{t-1})a_{ij}}{\sum_{i=1}^{N} \phi_{t-1}(i)f_i(\mathbf{x}_{t-1})}
\end{aligned} \tag{32}$$

5.2 Le "forward-backward" : Lisseur d'état

Si l'on s'intéresse à des quantités *lissées* du type $\gamma_t(i) \triangleq P(S_t = i|\mathbf{x}_{1:T})$ avec $t \leq T$ la situation est un peu plus complexe du fait du conditionnement à la fois par rapport à des observations passées et future (par rapport à l'indice courant t). Au début des années 1970 E. Baum et ses collègues ont proposé une solution à se problème qui est connu sous le nom d'algorithme de Baum-Welch, ou *forward-backward*. Cette dernière dénomination donne exactement le principe de la méthode : pour estimer les quantités lissées il suffit de combiner un prédicteur avant (en fonction du passé du processus depuis $\mathbf{X}_1$) et un prédicteur arrière (en fonction du futur du processus jusqu'à $\mathbf{X}_T$). Le prédicteur avant se calculant par une procédure de récursion dans le sens des indices t croissants (de 1 à T, ou *forward*), et le prédicteur arrière par une recursion pour les indices de T à 1 (*backward*).

Plus précisement, on définit $\alpha_t(i) \triangleq p(\mathbf{x}_{1:t}, S_t = i)$ et $\beta_t(i) \triangleq p(\mathbf{x}_{t+1:T}|S_t = i)$ les variables forward et backward, et on montre les relations suivantes

$$\boxed{\begin{cases} \alpha_1(i) &= f_i(\mathbf{x}_1)\pi_i \\ \alpha_{t+1}(j) &= f_j(\mathbf{x}_{t+1})\sum_{i=1}^{M} \alpha_t(i)a_{ij} \end{cases}} \tag{33}$$

$$\boxed{\left\{\begin{array}{rcl}\beta_T(i) & = & 1 \\ \beta_t(i) & = & \displaystyle\sum_{j=1}^{M} a_{ij} f_j(\mathbf{x}_{t+1})\beta_{t+1}(j)\end{array}\right.} \tag{34}$$

et

$$\boxed{\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\displaystyle\sum_{j=1}^{M} \alpha_t(j)\beta_t(j)}} \tag{35}$$

Exercice : Vérifier les relations (33) et (34) en procédant comme dans (32), c'est à dire en se ramenant aux hypothèses du modèle (25) et (26), éventuellement grâce à (30).

On montre de la même manière que la procédure de forward backward permet de calculer la quantité $\xi_t(i,j) \triangleq P(S_t = i, S_{t+1} = j|\mathbf{x}_{1:T})$ grâce à la relation [11]

$$\boxed{\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}f_j(\mathbf{x}_{t+1})\beta_{t+1}(j)}{\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)a_{ij}f_j(\mathbf{x}_{t+1})\beta_{t+1}(j)}} \tag{36}$$

Lors du calcul des variables forward et backward, il est fréquent voire inévitable de rencontrer des problèmes de dépassement des possibilités de représentation numérique (*overflows* ou *underflows*). En effet, les relations (33) et (34) montrent que les variables $\alpha_t(i)$ et $\beta_t(i)$ ont une dynamique qui tend à croître de manière exponentielle avec t (ou $T-t$ pour $\beta_t(i)$). [10] propose une solution à se problème qui consiste à normaliser les variables forward backward (procédure dite de *scaling*). Pour la variable forward, on utilise la relation (33) suivie, à chaque étape, d'une normalisation par $c_t = \sum_{i=1}^{N} \alpha_t(i)$ des $\alpha_t(i)$. Bien sûr cette normalisation modifie la variable backward (par exemple, au lieu de calculer $\alpha_T(i)$ on obtient $\alpha_T(i)/(\prod_{t=1}^{T} c_t)$) mais elle ne remet pas en cause la validité de la relation (35). La solution suggérée par [10] consiste à utiliser les mêmes facteurs c_t pour normaliser $\beta_t(i)$.

6 Algorithme EM pour les HMM

6.1 Formules de réestimation de base

En procédant comme au paragraphe 3.1, on écrit la quantité intermédiaire de l'algorithme EM sous la forme

$$\begin{aligned} Q_{\boldsymbol{\Theta}_n}(\boldsymbol{\Theta}) &= E\left[\log p(\mathbf{x}_{1:T}, s_{1:T}; \boldsymbol{\Theta})|\, \mathbf{x}_{1:T}; \boldsymbol{\Theta}_n\right] \\ &= E\left[\log p(\mathbf{x}_{1:T}|s_{1:T}; \boldsymbol{\Theta})|\, \mathbf{x}_{1:T}; \boldsymbol{\Theta}_n\right] + E\left[\log p(s_{1:T}; \boldsymbol{\Theta})|\, \mathbf{x}_{1:T}; \boldsymbol{\Theta}_n\right] \end{aligned} \tag{37}$$

Soit, en introduisant les données non observables

$$\begin{aligned} Q_{\boldsymbol{\Theta}_n}(\boldsymbol{\Theta}) = &\sum_{t=1}^{T}\sum_{i=1}^{N} \log f_i(\mathbf{x}_t; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) E[\mathbb{I}_{(S_t=i)}|\mathbf{x}_{1:T}; \boldsymbol{\Theta}_n] \\ &+ \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} \log a_{ij} E[\mathbb{I}_{(S_t=i,S_{t+1}=j)}|\mathbf{x}_{1:T}; \boldsymbol{\Theta}_n] + \sum_{i=1}^{N} \log \pi_i E[\mathbb{I}_{(S_1=i)}|\mathbf{x}_{1:T}; \boldsymbol{\Theta}_n] \end{aligned}$$

$$Q_{\boldsymbol{\Theta}_n}(\boldsymbol{\Theta}) = \sum_{t=1}^{T}\sum_{i=1}^{N} \gamma_t(i;\boldsymbol{\Theta}_n) \log f_i(\mathbf{x}_t;\boldsymbol{\mu_i},\boldsymbol{\Sigma_i}) + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} \xi_t(i,j;\boldsymbol{\Theta}_n)\log a_{ij} + \sum_{i=1}^{N}\gamma_1(i;\boldsymbol{\Theta}_n)\log\pi_i \quad (38)$$

La forme de la quantité intermédiaire de l'algorithme EM est donc très proche de celle obtenue pour le modèle de mélange en (19). Elle montre par ailleurs que l'étape E consiste simplement à calculer les probabilités conditionnelles $\gamma_t(i;\boldsymbol{\Theta}_n)$ et $\xi_t(i,j;\boldsymbol{\Theta}_n)$ en utilisant la procédure forward backward décrite précédemment. Comme dans le cas du mélange ce calcul se fait en considérant la valeur courante $\boldsymbol{\Theta}_n$ des paramètres du modèle (et comme dans le cas du mélange, on note dans la suite la dépendance en $\boldsymbol{\Theta}_n$ par un exposant (n) pour plus de lisibilité).

La partie maximisation par rapport aux différents paramètres du modèles fait intervenir trois quantités différentes. Pour les paramètres des densités gaussiennes, c'est le premier terme de (38) qui intervient. Dans la mesure où celui-ci est exactement l'analogue de celui obtenu pour le mélange en (19), les équations de réestimation sont donc exactement identiques à (23)-(24). Pour la distribution initiale $\boldsymbol{\pi}$, c'est le troisième terme de (19) qu'il est nécessaire de considérer. Compte tenu de la contrainte $\sum_{i=1}^{N}\pi_i = 1$, la résolution se fait comme dans le cas du modèle de mélange, et on obtient une formule analogue à (22). Enfin, pour la matrice de transition, chaque ligne vérifie une contrainte de normalisation qui conduit à introduire autant de multiplicateurs de Lagrange. En procédant comme dans le cas des mélanges (pour les poids du mélange), on trouve

$$\boxed{a_{ij}^{(n+1)} = \frac{\sum_{t=1}^{T-1}\xi_t^{(n)}(i,j)}{\sum_{j=1}^{N}\sum_{t=1}^{T-1}\xi_t^{(n)}(i,j)} = \frac{\sum_{t=1}^{T-1}\xi_t^{(n)}(i,j)}{\sum_{t=1}^{T-1}\gamma_t^{(n)}(i)}} \quad (39)$$

6.2 Paramètres liés

Pour le traitement de la parole, il est souvent nécessaire de simplifier la structure du modèle afin de limiter la quantité de données d'apprentissage nécessaire, qui est en général déjà considérable. Ainsi, on utilise en pratique systématiquement des matrices de covariances diagonales $\boldsymbol{\Sigma}_i$. Une autre solution très utilisée consiste à partager certains paramètres entre plusieurs états du modèle [13]. Les modifications à apporter aux équations de réestimation ci-dessus sont en général immédiates dans la mesure où il suffit de tenir compte de ces contraintes lorsque l'on différencie la quantité intermédiaire de l'algorithme EM.

Pour fixer les idées, supposons par exemple que chaque état soit caractérisé par un vecteur moyen $\boldsymbol{\mu}_i$ mais que par contre, tous les états partagent la même matrice de covariance $\boldsymbol{\Sigma}$. On vérifie dans ce cas que la formule de réestimation des vecteurs moyen est inchangée et que la différentielle de la quantité intermédiaire de l'EM par rapport à $\boldsymbol{\Phi} = \boldsymbol{\Sigma}^{-1}$ s'écrit

$$-\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\gamma_t^{(n)}(i)\left(-\boldsymbol{\Phi}^{-1} + (\mathbf{x}_t - \boldsymbol{\mu}_i)(\mathbf{x}_t - \boldsymbol{\mu}_i)'\right)$$

L'optimum étant donc obtenu pour

$$\boldsymbol{\Sigma}^{(n+1)} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{N}\gamma_t^{(n)}(i)(\mathbf{x}_t - \boldsymbol{\mu}_i^{(n+1)})(\mathbf{x}_t - \boldsymbol{\mu}_i^{(n+1)})'}{T} \tag{40}$$

Cette dernière équation peut d'ailleurs se réécrire de manière équivalente en faisant intervenir les quantités

$$\boldsymbol{\Sigma}_i^{(n+1)} = \frac{\sum_{t=1}^{T}\gamma_t^{(n)}(i)(\mathbf{x}_t - \boldsymbol{\mu}_i^{(n+1)})(\mathbf{x}_t - \boldsymbol{\mu}_i^{(n+1)})'}{\sum_{t=1}^{T}\gamma_t^{(n)}(i)}$$

c'est à dire les valeurs estimées dans le cas où les matrices de covariance de chaque état sont distinctes. On obtient

$$\boldsymbol{\Sigma}^{(n+1)} = \frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\gamma_t^{(n)}(i)\right)\boldsymbol{\Sigma}_i^{(n+1)}}{T} \tag{41}$$

6.3 Séquences d'apprentissage multiples

En traitement de la parole, on s'intéresse en général à des HMM dit gauche-droite pour lesquels la matrice de transition est triangulaire supérieure. Dans un tel modèle, S_t visite successivement les états de 1 à N, en sautant éventuellement des états, mais sans retour en arrière possible. Pour estimer efficacement les paramètres d'un modèle gauche-droite, il est nécessaire de disposer de plusieurs séquence d'observations. Si l'on suppose que les différentes séquences d'observations sont indépendantes, la modification à apporter à la procédure d'estimation est relativement directe dans la mesure où la quantité intermédiaire de l'EM sera une somme des contributions propres au différentes séquences. L'étape E doit être effectuée indépendamment pour chacune des séquences d'apprentissage puisque celle-ci sont indépendantes. Pour l'étape M, il est nécessaire de maximiser une somme de termes et on trouve des formules de réestimation du type [10]

$$\boldsymbol{\mu}_i^{(n+1)} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_t^{(r,n)}(i)\mathbf{x}_{r,t}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_t^{(r,n)}(i)} \tag{42}$$

pour la moyenne $\boldsymbol{\mu}_i$, où R désigne le nombre de séquences d'apprentissage, $\mathbf{x}_{r,1}, \ldots, \mathbf{x}_{r,T_r}$ les données correspondant à chaque séquence, et $\gamma_1^{(r,n)}(i)$ à $\gamma_{T_r}^{(r,n)}(i)$ les probabilités a posteriori estimées pour cette séquence.

6.4 Apprentissage sous-optimal

Dans beaucoup de systèmes de reconnaissance de parole, on utilise pour accélérer l'apprentissage un critère de vraisemblance réduite en cherchant à maximiser la quantité

$$g(\boldsymbol{\Theta}) = \max_{s_{1:T}} p(s_{1:T}, \mathbf{x}_{1:T}; \boldsymbol{\Theta})$$

c'est à dire la vraisemblance complète optimisée vis à vis de la séquence d'états inconnus (méthode dite d'apprentissage *segmental* en anglais, dans la mesure où elle fait appel à une segmentation déterministe des observations) [11], [7]. Cette quantité est maximisée itérativement par un algorithme dit de relaxation cyclique [9] (maximisations partielles alternées selon $s_{1:T}$ et $\mathbf{\Theta}$) qui ressemble formellement à l'EM :

1. Estimation de la séquence optimale $s_{1:T}^{(n)}$ étant donné les observations et les paramètres courants $\mathbf{\Theta}_n$ par l'algorithme de programmation dynamique (ou "de Viterbi") [11], [10].

2. Maximisation de la vraisemblance complète avec substitution de la séquence d'états inconnue par la séquence déterminée ci-dessus. Formellement, ceci revient à utiliser les formules de réestimation de l'EM avec un vecteur de probabilité dégénéré pour lequel $\gamma_t(s_t^{(n)}) = 1$ (et $\gamma_t(i) = 0$ pour $i \neq s_t^{(n)}$).

D'un point de vue pratique cette méthode est plus rapide parce qu'elle évite une sommation sur tous les états possibles lors de la réestimation des paramètres. Bien que cette méthode ne présente pas les garanties d'optimalité du maximum de vraisemblance, elle semble fournir de bons résultats en pratique pour le type HMM utilisés en traitement de la parole.

References

[1] T. W. Anderson. *An introduction to multivariate statistical analysis.* J. Wiley, New York, 1958.

[2] M. Charbit and E. Moulines. *Cours d'estimation statistique.* ENST, département Signal, 1998.

[3] C. Chatfield and A. J. Collins. *Introduction to multivariate analysis.* Chapman and Hall, 1980.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, 39(1):1–38 (with discussion), 1977.

[5] R. A. Horn and C. R. Johnson. *Matrix analysis.* Cambridge University Press, 1985.

[6] R. A. Horn and C. R. Johnson. *Topics in matrix analysis.* Cambridge University Press, 1991.

[7] B-H. Juang and L. R. Rabiner. The segmental k-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust., Speech, Signal Processing*, 38(9):1639–1641, September 1990.

[8] I. L. MacDonald and W. Zucchini. *Hidden Markov models and other models for discrete-valued time series.* Chapman & Hall, 1997.

[9] M. Minoux. *Programmation Mathématique : théorie et algorithmes (2 tomes).* Collection technique et scientifique des télécommunications. Dunod, 1983.

[10] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–285, February 1989.

[11] L. R. Rabiner and B-H. Juang. *Fundamentals of speech recognition*. Prentice-Hall, 1993.

[12] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley, 1985.

[13] S. Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, 13(5), September 1996.