Introduction au Traitement Automatique de la Parole

Notes de cours / DEC2

Première édition

Faculté Polytechnique de Mons

Thierry Dutoit
Faculté Polytechnique de Mons
TCTS Lab
Ave. Copernic Ph: +32 65 374774
Parc Initialis Fax: +32 65 374729
B-7000 Mons Thierry.Dutoit@fpms.ac.be
Belgium http://tcts.fpms.ac.be/~dutoit

AVANT-PROPOS

Ce document reprend un ensemble de notions de base enseignées aux étudiants du Diplôme d'Etudes Complémentaires Interuniversitaire en Ingénierie de la Langue (DEC2) organisé conjointement par les Facultés Universitaires Notre-Dame de la Paix à Namur, l'Université Catholique de Louvain, l'Université de Liège, et la Faculté Polytechnique de Mons (voir le site web du DEC2 : http://www.info.fundp.ac.be/~gde/dec.html), dans le cadre du cours d'Introduction au Traitement de la Parole (voir le site web du cours : http://tcts.fpms.ac.be/cours/1005-08/speech/).

Les étudiant(e)s abordant ce cours ont préalablement suivi un cours de mise a niveau en mathématiques et en traitement du signal, qui leur permet de mieux comprendre les notions d'analyse spectrale et de filtrage numérique. Il n'est cependant pas question ici d'entrer dans le détail mathématique des principes et algorithmes utilisés en traitement de parole, mais plutôt d'esquisser les principes généraux, et de donner un aperçu de l'état de l'art.

On trouvera dans ce document une introduction générale à l' « objet parole » dont il sera question dans la suite du cours, un exposé introductif aux modèles mathématiques utilisés pour le traitement automatique de cet objet, et un aperçu des problèmes posés par le codage, la synthèse, et la reconnaissance de la parole, ainsi que des solutions qu'on peut aujourd'hui y apporter.

Une partie du texte et des images qui constituent ces notes de cours a servi de canevas à certains chapitres de l'ouvrage de référence suivant :

> ***Traitement de la Parole***, R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich, Presses Polytechniques Universitaires Romandes, Lausanne, 2000.

Nous conseillons vivement au lecteur intéressé de s'y reporter pour plus de détail et de précision.

T. Dutoit, Mons, le 20 octobre 2000

CHAPITRE 1
INTRODUCTION

1.1 Le traitement de la parole

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications.

L'importance particulière du traitement de la parole dans ce cadre plus général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine.

L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, tient sans aucun doute au rôle fascinant que joue le cerveau humain à la fois dans la production et dans la compréhension de la parole et à l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée.

Pour mieux comprendre cette particularité, penchons-nous un instant sur d'autres vecteurs d'information. L'image, par exemple, n'existe que dans la mesure où elle est appelée à être perçue par l'œil, et, bien au-delà, interprétée par le cerveau. Les techniques de traitement de l'image pourront en tirer parti en prenant en compte, d'une part, les caractéristiques physiques de l'œil et, d'autre part, les propriétés perceptuelles que lui confère le cortex visuel. Un exemple bien connu de ce type d'influence du récepteur sur le mode de traitement des signaux associés nous est fourni par l'image vidéo, dont les 24 images/seconde découlent directement du phénomène de persistance rétinienne. A l'inverse, un signal d'origine biologique tel que l'électro-myogramme, qui mesure l'état d'activité d'un muscle, n'existe que dans la mesure où il est produit par ce muscle, sous le contrôle étroit du cortex moteur. Une bonne connaissance du muscle sera par conséquent un pré-requis indispensable au traitement automatique de l'électro-myogramme correspondant.

Aucun de ces signaux, pourtant fort complexes, n'est cependant à la fois appelé à être *produit* et *perçu* instantanément par le cerveau, comme c'est le cas pour la parole. La parole est en effet produite par le conduit vocal, contrôlé en permanence par le cortex moteur. L'étude des mécanismes de phonation permettra donc de déterminer, dans une certaine mesure, ce qui est parole et ce qui n'en est pas. De même, l'étude des mécanismes d'audition et des propriétés perceptuelles qui s'y rattachent permettra de dire ce qui, dans le signal de parole, est réellement perçu. Mais l'essence même du signal de parole ne peut être cernée de façon réaliste que dans la mesure où l'on imagine, bien au-delà de la simple mise en commun des propriétés de production et de perception de la parole, les propriétés du signal dues à la mise en boucle de ces deux fonctions. Mieux encore, c'est non seulement la perception de la parole qui vient influer sur sa production par le biais de ce bouclage, mais aussi et surtout sa *compréhension*[1]. On ne parle que dans la mesure ou l'on s'entend et où l'on se comprend soi-même; la complexité du signal qui en résulte s'en ressent forcément[2].

S'il n'est pas en principe de parole sans cerveau humain pour la produire, l'entendre, et la comprendre, les techniques modernes de traitement de la parole tendent cependant à produire des systèmes automatiques qui se substituent à l'une ou l'autre de ces fonctions :

- Les ***analyseurs*** de parole cherchent à mettre en évidence les caractéristiques du signal vocal tel qu'il est produit, ou parfois tel qu'il est perçu (on parle alors d'*analyseur perceptuel*), mais jamais tel qu'il est compris, ce rôle étant réservé aux reconnaisseurs. Les analyseurs sont utilisés soit comme composant de base de systèmes de codage, de reconnaissance ou de synthèse (voir ci-dessous), soit en tant que tels pour des applications spécialisées, comme l'aide au diagnostic médical (pour les pathologies du larynx, par analyse du signal vocal) ou l'étude des langues.
- Les ***reconnaisseurs*** ont pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse. On distingue fondamentalement deux types de reconnaissance, en fonction de l'information que l'on cherche à extraire du signal vocal : la ***reconnaissance du locuteur***, dont l'objectif est de reconnaître la personne qui parle, et la ***reconnaissance de la parole***, où l'on s'attache plutôt à reconnaître ce qui est dit. On classe également les reconnaisseurs en fonction des hypothèses simplificatrices sous lesquelles ils sont appelés à fonctionner. Ainsi :
    - En reconnaissance du locuteur, on fait la différence entre l'***identification*** et la ***vérification*** du locuteur, selon que le problème

---

1 Ce qui accroît encore la différence entre la parole et, par exemple, l'image : alors que la compréhension de l'image est en principe également accessible à tous, la compréhension de la parole est le résultat d'un apprentissage socio-culturel lié à une *communauté linguistique*.

2 A cet égard, la discipline scientifique qui s'apparente le plus au traitement de la parole est sans doute le traitement automatique des caractères manuscrits.

est de vérifier que la voix analysée correspond bien à la personne qui est sensée la produire, ou qu'il s'agit de déterminer qui, parmi un nombre fini et préétabli de locuteurs, a produit le signal analysé.

  - On sépare reconnaissance du locuteur ***dépendante du texte***, reconnaissance ***avec texte dicté***, et reconnaissance ***indépendante du texte***. Dans le premier cas, la phrase à prononcer pour être reconnu est fixée dès la conception du système; elle est fixée lors du test dans le deuxième cas, et n'est pas précisée dans le troisième.
  - On parle de reconnaisseur de parole ***monolocuteur***, ***multilocuteur***, ou ***indépendant du locuteur***, selon qu'il a été entraîné à reconnaître la voix d'une personne, d'un groupe fini de personnes, ou qu'il est en principe capable de reconnaître n'importe qui.
  - On distingue enfin reconnaisseur de ***mots isolés***, reconnaisseur de ***mots connectés***, et reconnaisseur de ***parole continue***, selon que le locuteur sépare chaque mot par un silence, qu'il prononce de façon continue une suite de mots prédéfinis, ou qu'il prononce n'importe quelle suite de mots de façon continue.

- Les ***synthétiseurs*** ont quant à eux la fonction inverse de celle des analyseurs et des reconnaisseurs de parole : ils produisent de la parole artificielle. On distingue fondamentalement deux types de synthétiseurs : les ***synthétiseurs de parole à partir d'une représentation numérique***, inverses des analyseurs, dont la mission est de produire de la parole à partir des caractéristiques numériques d'un signal vocal telles qu'obtenues par analyse, et les ***synthétiseurs de parole à partir d'une représentation symbolique***¸ inverse des reconnaisseurs de parole et capables en principe de prononcer n'importe quelle phrase sans qu'il soit nécessaire de la faire prononcer par un locuteur humain au préalable. Dans cette seconde catégorie, on classe également les synthétiseurs en fonction de leur mode opératoire :
  - Les ***synthétiseurs à partir du texte*** reçoivent en entrée un texte orthographique et doivent en donner lecture.
  - Les ***synthétiseurs à partir de concepts***, appelés à être insérés dans des systèmes de dialogue homme-machine, reçoivent le texte à prononcer et sa structure linguistique, telle que produite par le système de dialogue.
- Enfin, le rôle des ***codeurs*** est de permettre la transmission ou le stockage de parole avec un débit réduit, ce qui passe tout naturellement par une prise en compte judicieuse des propriétés de production et de perception de la parole.

On comprend aisément que, pour obtenir de bons résultats dans chacune de ces tâches, il faut tenir compte des caractéristiques du signal étudié. Et, vu la complexité de ce signal, due en grande partie au couplage étroit entre production, perception, et compréhension, il n'est pas étonnant que les recherches menées par les spécialistes soient directement liées aux progrès obtenus dans de nombreuses autres disciplines scientifiques, progrès dont elles sont par ailleurs souvent à la fois les bénéficiaires et les instigatrices. Comme le

remarque très justement Allen: *"Le traitement de la parole fournit d'excellents exemples pour l'étude de systèmes complexes, dans la mesure où il soulève des questions fondamentales dans les domaines du partitionnement des systèmes, du choix d'unités descriptives, des techniques de représentation, des niveaux d'abstraction, des formalismes de représentation de la connaissance, de l'expression d'interactions entre contraintes, des techniques de modularité et de hiérarchisation, des techniques d 'estimation de vraisemblance, des techniques de mesure de la qualité et du naturel d'un stimulus, de la détermination de classes d'équivalence, de la paramétrisation de modèles adaptatifs, de l'étude des compromis entre représentations procédurales et déclaratives, de l'architecture des systèmes, et de l'exploitation des technologies modernes pour produire des systèmes qui fonctionnent en temps réel pour un coût acceptable*".

1.2 Qu'est-ce que la parole ?

L'information portée par le signal de parole peut être analysée de bien des façons. On en distingue généralement plusieurs niveaux de description non exclusifs : *acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique,* et *pragmatique*.

1.2.1 Le niveau acoustique

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La ***phonétique acoustique***[3] étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone (lui-même associé à un préamplificateur). De nos jours, le signal électrique résultant est le plus souvent numérisé. Il peut alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les ***traits acoustiques*** : sa ***fréquence fondamentale***, son ***énergie***, et son ***spectre***. Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : ***pitch***, ***intensité***, et ***timbre***.

L'opération de numérisation, schématisée à la figure 1.1, requiert successivement : un ***filtrage de garde***, un ***échantillonnage***, et une ***quantification***.

---

[3] Dans la suite, nous présentons les niveaux acoustiques et phonétiques comme s'ils étaient indépendants bien que, stricto sensu, les aspects acoustiques de la parole sont du ressort d'une branche particulière de la phonétique : la phonétique acoustique (les autres étant la phonétique physiologique ou articulatoire, et la phonétique perceptive). La phonétique articulatoire fait l'objet du paragraphe 1.2.2.

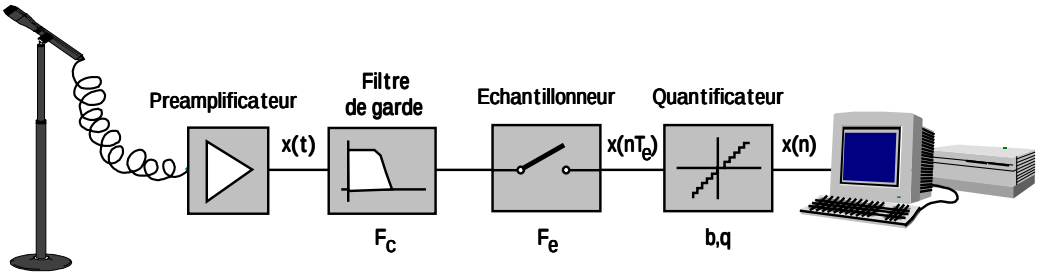


Fig. 1.1 Enregistrement numérique d'un signal acoustique. La fréquence de coupure du filtre de garde, la fréquence d'échantillonnage, le nombre de bits et le pas de quantification sont respectivement notés f_c , f_e , b, et q.

1.2.1.1 Audiogramme

L'échantillonnage transforme le signal à temps continu *x(t)* en signal à temps discret $x(nT_e)$ défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage T_e; celle-ci est elle-même l'inverse de la fréquence d'échantillonnage f_e. Pour ce qui concerne le signal vocal, le choix de f_e résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz. Il faut donc en principe choisir une fréquence f_e égale à 24 kHz au moins pour satisfaire raisonnablement au théorème de Shannon[4]. Cependant, le coût d'un traitement numérique, filtrage, transmission, ou simplement enregistrement peut être réduit d'une façon notable si l'on accepte une limitation du spectre par un filtrage préalable. C'est le rôle du filtre de garde, dont la fréquence de coupure f_c est choisie en fonction de la fréquence d'échantillonnage retenue. Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit f_e = 8000 Hz. Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole, la fréquence peut varier de 6000 à 16000 Hz. Par contre pour le signal audio (parole et musique), on exige une bonne représentation du signal jusque 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz. Pour les applications multimédia, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées : 22.5 kHz, 11.25 kHz.

Parmi le continuum des valeurs possibles pour les échantillons $x(nT_e)$, la quantification ne retient qu'un nombre fini *2b* de valeurs (*b* étant le nombre de bits de la quantification), espacées du pas de quantification *q*. Le signal numérique résultant est noté *x(n)*. La quantification produit une erreur de quantification qui normalement se comporte comme un bruit blanc; le pas de quantification est donc imposé par le rapport signal à bruit à garantir. Si le pas de quantification est constant, ce rapport est fonction de l'amplitude du signal; les signaux de faible amplitude sont dès lors mal représentés. Aussi adopte-t-on

---

4 Suivant le *théorème de Shannon,* les signaux doivent être échantillonnés à une fréquence d'échantillonnage supérieure ou égale à deux fois leur plus haute composante fréquentielle. Ils doivent être filtrés passe-bas dans le cas contraire.

pour la transmission téléphonique une loi de quantification logarithmique et chaque échantillon est représenté sur 8 bits (256 valeurs). Par contre, la quantification du signal musical exige en principe une quantification linéaire sur 16 bits (65536 valeurs).

Une caractéristique essentielle qui résulte du mode de représentation est le débit binaire, exprimé en bits par seconde (b/s), nécessaire pour une transmission ou un enregistrement du signal vocal. La transmission téléphonique classique exige un débit de 8 kHz x 8 bits = 64 kb/s; la transmission ou l'enregistrement d'un signal audio exige en principe un débit de l'ordre de 48 kHz x 16 bits = 768 kb/s (à multiplier par deux pour un signal stéréophonique)[5].

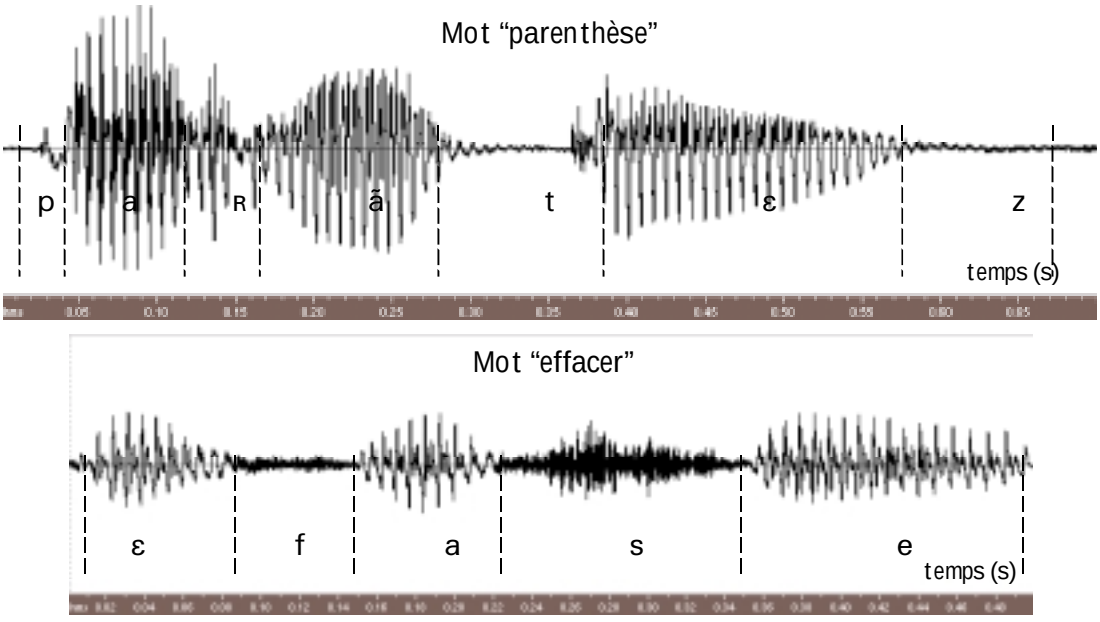


Fig. 1.2 Audiogramme de signaux de parole.

5 La redondance naturelle du signal vocal permet de réduire le débit binaire dans une très large mesure, au prix d'un traitement plus ou moins complexe et au risque d'une certaine dégradation de la qualité de la représentation. Cette question sera abordée en détail au chapitre 4, consacré au codage de la parole.

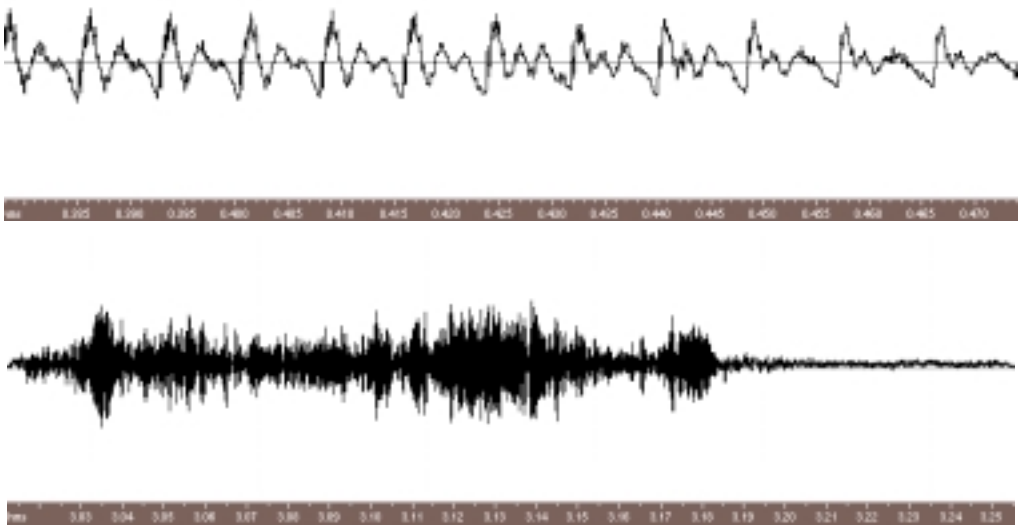


Fig. 1.3 Exemples de son voisé (haut) et non-voisé (bas).

La figure 1.2 représente l'évolution temporelle, ou *audiogramme*, du signal vocal pour les mots 'parenthèse', et 'effacer'. On y constate une alternance de zones assez périodiques et de zones bruitées, appelées zones ***voisées*** et ***non-voisées***. La figure 1.3 donne une représentation plus fine de tranches de signaux voisés et non voisés. L'évolution temporelle ne fournit cependant pas directement les traits acoustiques du signal. Il est nécessaire, pour les obtenir, de mener à bien un ensemble de calculs ad-hoc.

1.2.1.2 Transformée de Fourier à court terme

La transformée de Fourier à court terme est obtenue en extrayant de l'audiogramme une 30aine de ms de signal vocal, en pondérant ces échantillons par une fenêtre de pondération (souvent une fenêtre de Hamming) et en effectuant un transformée de Fourier sur ces échantillons.

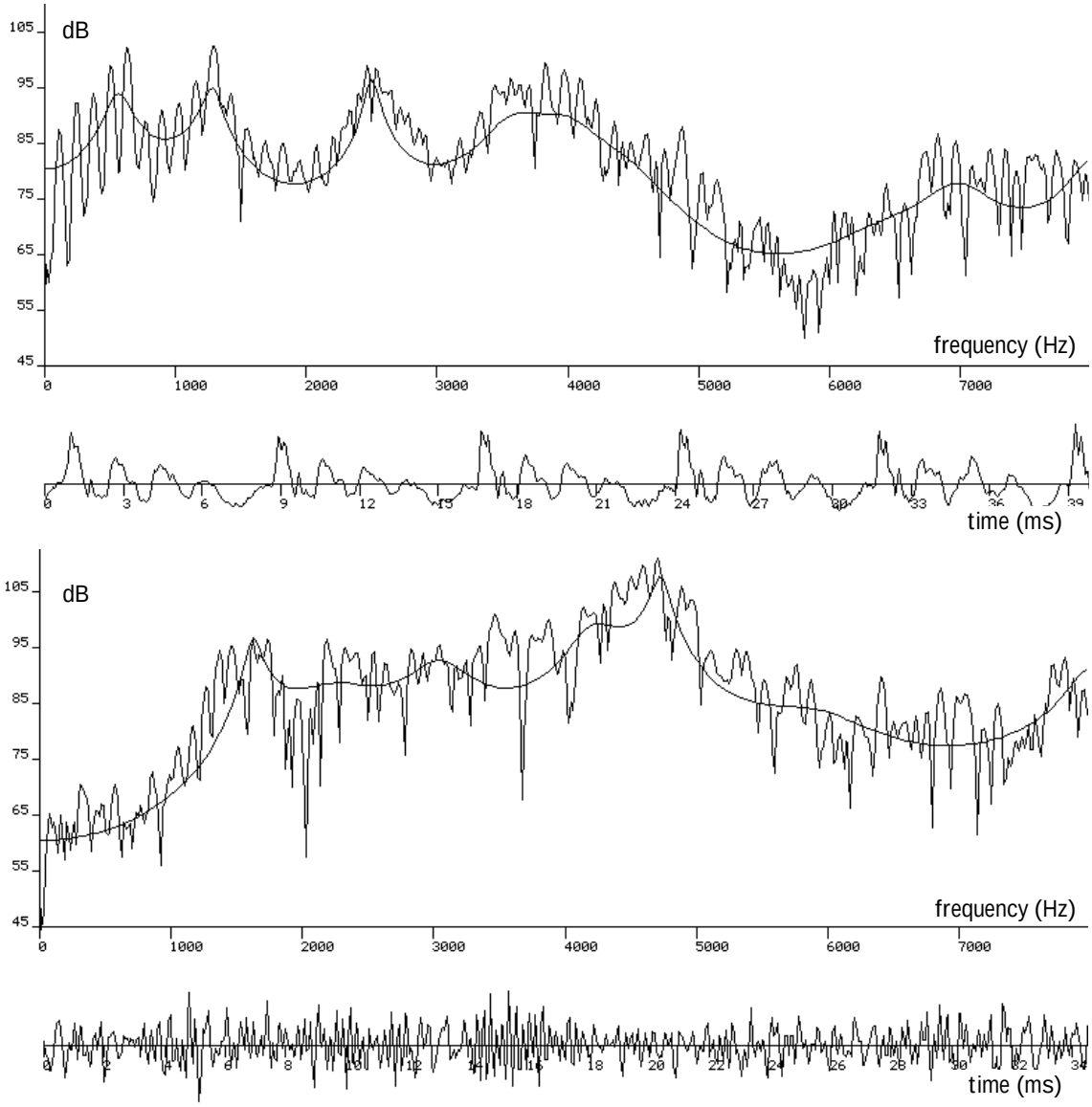


Fig. 1.9 Evolution temporelle (en haut) et transformée de Fourier discrète (en bas) du [a] et du [ʃ] de *'baluchon'* (signaux pondérés par une fenêtre de Hamming de 30 ms).

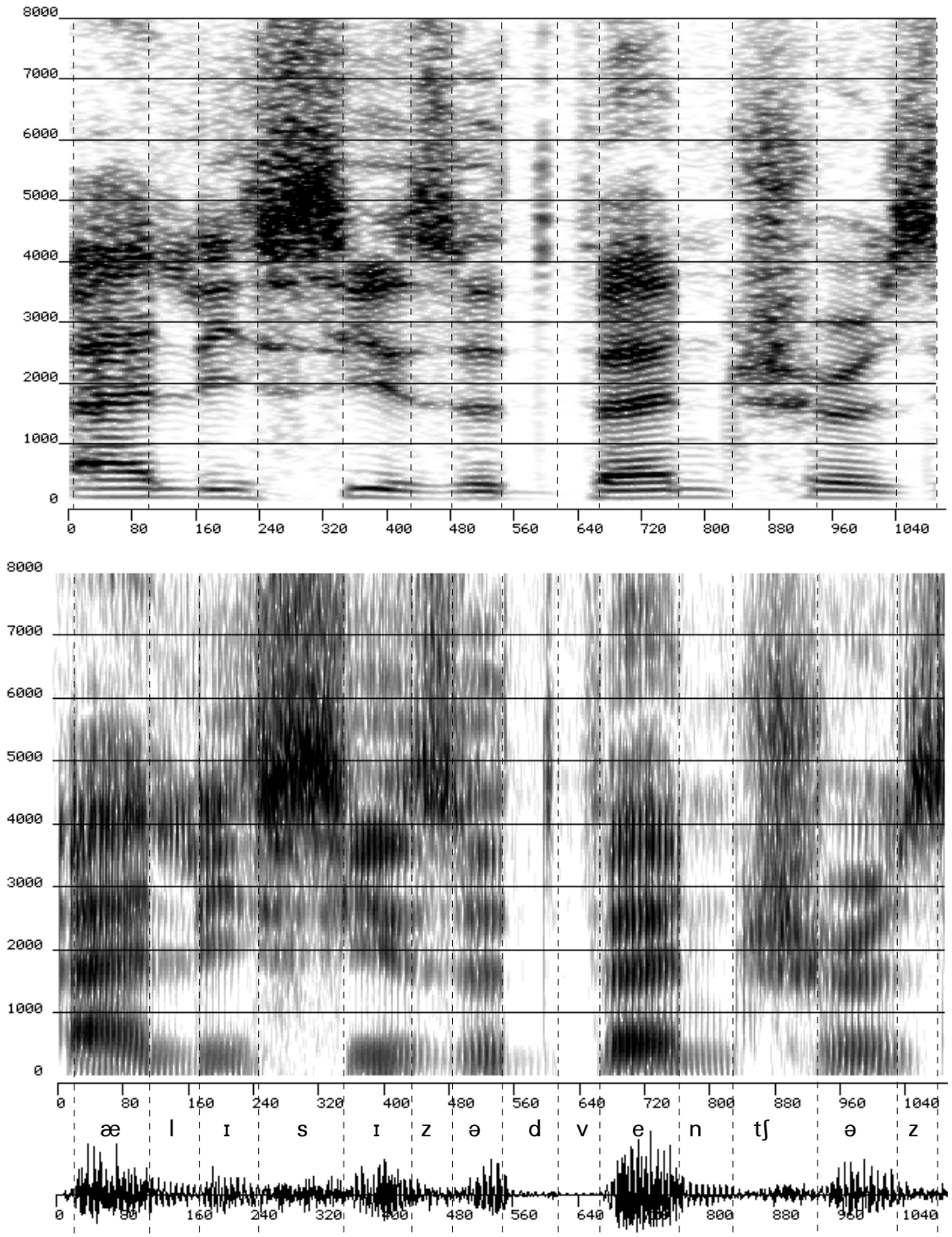


Fig. 1.10 Spectrogrammes à large bande (en bas), à bande étroite (en haut), et évolution temporelle de la phrase anglaise *'Alice's adventures'*, échantillonnée à 11.25 kHz (calcul avec fenêtres de Hamming de 10 et 30 ms respectivement).

La figure 1.9 illustre la transformée de Fourier d'une tranche voisée et celle d'une tranche non-voisée. Les parties voisées du signal apparaissant sous la forme de successions de pics spectraux marqués, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière. La forme générale de ces spectres, appelée ***enveloppe spectrale***, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal et sont appelés ***formants*** et ***anti-formants***. L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son. Il apparaît en pratique que l'enveloppe spectrale des sons voisés est de type passe-bas, avec environ un formant par kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante au timbre. Par contre, les sons non-voisés présentent souvent une accentuation vers les hautes fréquences.

1.2.1.3 Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un ***spectrogramme***. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de spectrogramme ***à large bande*** ou ***à bande étroite*** selon la durée de la fenêtre de pondération (Fig. 1.10). Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms); ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales.

1.2.1.4 Fréquence fondamentale

Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la fréquence fondamentale[6] ou *pitch*.

La figure 1.8 donne l 'évolution temporelle de la fréquence fondamentale de la phrase *"les techniques de traitement de la parole"*. On constate qu'à l'intérieur des zones voisées la fréquence fondamentale évolue lentement dans le temps. Elle s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants.

---

6 Nous les aborderons plus loin (chapitre 4).

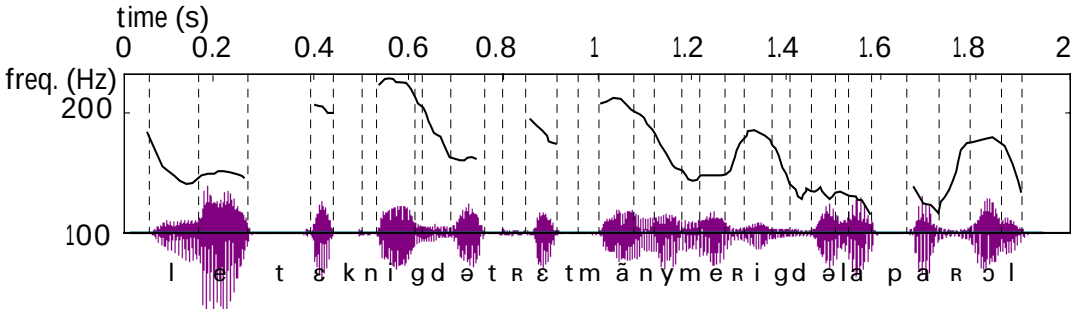


Fig. 1.8 Evolution de la fréquence de vibration des cordes vocales dans la phrase "*les techniques de traitement numérique de la parole*". La fréquence est donnée sur une échelle logarithmique; les sons non-voisés sont associés à une fréquence nulle..

1.2.2 Le niveau phonétique

Au contraire des acousticiens, ce n'est pas tant le signal qui intéresse les phonéticiens que la façon dont il est produit par le système articulatoire, présenté à la figure 1.11, et perçu par le système auditif.

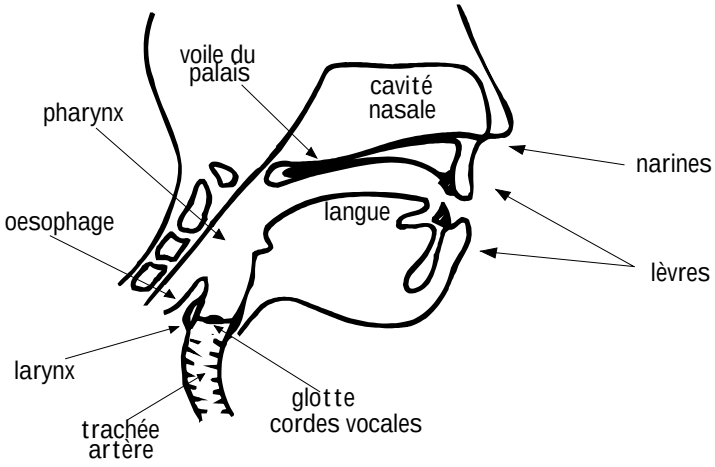


Fig. 1.11 L'appareil phonatoire.

1.2.2.1 Phonation

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations kinesthésiques.

L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le ***larynx*** ou la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui

entourent une cavité située à la partie supérieure de la trachée (Fig. 1.12). Les ***cordes vocales*** sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée ***glotte***. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés (ou ***sourds***[7]). Les sons voisés (ou ***sonores***) résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, et les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des cavités pharyngienne et buccale pour la plupart des sons. Lorsque la ***luette*** est en position basse, la cavité nasale vient s'y ajouter en dérivation. Notons pour terminer le rôle prépondérant de la langue dans le processus phonatoire. Sa hauteur détermine la hauteur du pharynx : plus la langue est basse, plus le pharynx est court. Elle détermine aussi le ***lieu d'articulation***, région de rétrécissement maximal du canal buccal, ainsi que l'***aperture***, écartement des organes au point d'articulation.

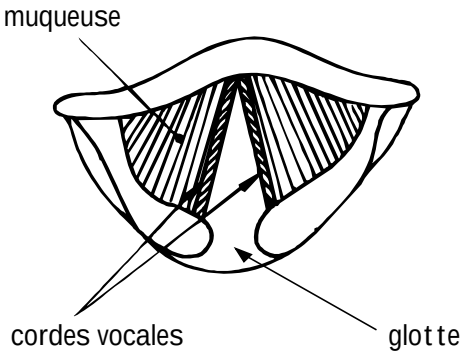


Fig. 1.12 Section du larynx, vu de haut.

1.2.2.2 L'alphabet phonétique international

L'***alphabet phonétique international*** (IPA) associe des symboles phonétiques aux sons, de façon à permettre l'écriture compacte et universelle des prononciations (voir tableau 1.2 pour le français).

---

7 Les phonéticiens appellent *sourd* ou *sonore* ce que les ingénieurs qualifient de *voisé* ou *non-voisé*.

Tableau 1.2 Les symboles de l'alphabet phonétique international utilisés en français.

IPA	EXEMPLES	IPA	EXEMPLES
i	idée, ami	p	patte, repas, cap
e	ému, ôté	t	tête, ôter, net
ɛ	perdu, modèle	k	carte, écaille, bec
a	alarme, patte	b	bête, habile, robe
ɑ	bâton, pâte	d	dire, rondeur, chaud
ɔ	Obstacle, corps	g	gauche, égal, bague
o	auditeur, beau	f	feu, affiche, chef
u	coupable, loup	s	sœur, assez, passe
y	punir, élu	ʃ	chanter, machine, poche
ø	creuser, deux	v	vent, inventer, rêve
œ	malheureux, peur	z	zéro, raisonner, rose
ə	petite, fortement	ʒ	jardin, manger, piège
ɛ̃	peinture, matin	l	long, élire, bal
ɑ̃	vantardise, temps	ʀ	rond, chariot, sentir
ɔ̃	rondeur, bon	m	madame, aimer, pomme
œ̃	lundi, brun	n	nous, punir, bonne
j	piétiner, briller		agneau, peigner, règne
w	oui, fouine	ŋ	jumping, smoking
ɥ	huile, nuire	h	halte, hop (exclamations)

1.2.2.3 Phonétique articulatoire

Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur ***mode articulatoire***. On distingue généralement trois classes principales : les ***voyelles***, les ***semi-voyelles*** et les ***liquides***, et les ***consonnes***.

Les voyelles [i,e,ɛ,ɑ,a,ɔ,o,y,u,ø,œ,ə,ɛ̃,ɑ̃,ɔ̃,œ̃] diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal (et non, comme on l'entend souvent dire, par le degré d'activité des cordes vocales, déjà mentionné sous le terme de ***voisement***). Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la bouche se réduit alors à une modification du timbre vocalique. Si, au contraire, le passage se rétrécit par endroit, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. La bouche est dans ce cas un organe de production à part entière. Les semi-voyelles [j,w,], quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes. Enfin, les liquides [l,ʀ] sont assez difficiles à classer. L'articulation de [l] ressemble à celle d'une voyelle, mais la

position de la langue conduit à une fermeture partielle du conduit vocal. Le son [ʀ], quant à lui, admet plusieurs réalisations fort différentes.

Les voyelles se différencient principalement les unes des autres par leur *lieu d'articulation,* leur *aperture*, et leur *nasalisation*. On distingue ainsi, selon la localisation de la masse de la langue, les voyelles ***antérieures***, les voyelles ***moyennes***, et les voyelles ***postérieures***, et, selon l'écartement entre l'organe et le lieu d'articulation, les voyelles ***fermées*** et ***ouvertes***. Les voyelles ***nasales*** [ɛ̃,ɑ̃,ɔ̃,œ̃] diffèrent des voyelles *orales* [i,e,ɛ,ɑ,a,ɔ,o,ɣ,u,ø,œ,ə] en ceci que le voile du palais est abaissé pour leur prononciation, ce qui met en parallèle les cavités nasales et buccale. Notons que, dans un contexte plus général que celui de la seule langue française, d'autres critères peuvent être nécessaires pour différencier les voyelles, comme leur *labialisation*, leur *durée*, leur *tension*, leur *stabilité*, leur *glottalisation*¸ voire même la *direction du mouvement de l'air*.

On classe principalement les consonnes en fonction de leur *mode d'articulation*, de leur *lieu d'articulation*, et de leur *nasalisation*. Comme pour les voyelles, d'autres critères de différenciation peuvent être nécessaires dans un contexte plus général : l'*organe articulatoire*, la *source sonore*, l'*intensité*, l'*aspiration*, la *palatalisation*, et la *direction du mouvement de l'air*.

En français, la distinction de mode d'articulation conduit à deux classes : les ***fricatives*** (ou ***constrictives***) et les ***occlusives*** (ou ***plosives***). Les fricatives sont créées par une constriction du conduit vocal au niveau du lieu d'articulation, qui peut être le palais [ʃ,ʒ], les dents [s,z], ou les lèvres [f,v]. Les fricatives non-voisées sont caractérisées par un écoulement d'air turbulent à travers la glotte, tandis que les fricatives voisées combinent des composantes d'excitation périodique et turbulente : les cordes vocales s'ouvrent et se ferment périodiquement, mais la fermeture n'est jamais complète. Les occlusives correspondent quant à elles à des sons essentiellement dynamiques. Une forte pression est créée en amont d'une occlusion maintenue en un certain point du conduit vocal (qui peut ici aussi être le palais [k,g], les dents [t,d], ou les lèvres [p,b]), puis relâché brusquement. La période d'occlusion est appelée la phase de tenue. Pour les occlusives voisées [b,d,g] un son basse fréquence est émis par vibration des cordes vocales pendant la phase de tenue; pour les occlusives non voisées [p,t,k], la tenue est un silence.

Enfin, les consonnes nasales [m,n, ,ŋ] font intervenir les cavités nasales par abaissement du voile du palais.

Les traits acoustiques du signal de parole sont évidemment liés à sa production. L'intensité du son est liée à la pression de l'air en amont du larynx. Sa fréquence, qui n'est rien d'autre que la fréquence du cycle d'ouverture/fermeture des cordes vocales, est déterminée par la tension de muscles qui les contrôlent. Son spectre résulte du filtrage dynamique du signal glottique (impulsions, bruit, ou combinaison des deux) par le conduit vocal, qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses. Ainsi, par exemple, on peut approximativement représenter les voyelles dans le plan des deux premiers formants (Fig. 1.13). On observe en pratique un certain recouvrement dans les zones formantiques correspondant à chaque voyelle (un

affichage en trois dimensions figurant les trois premiers formants permettrait une meilleure séparation).

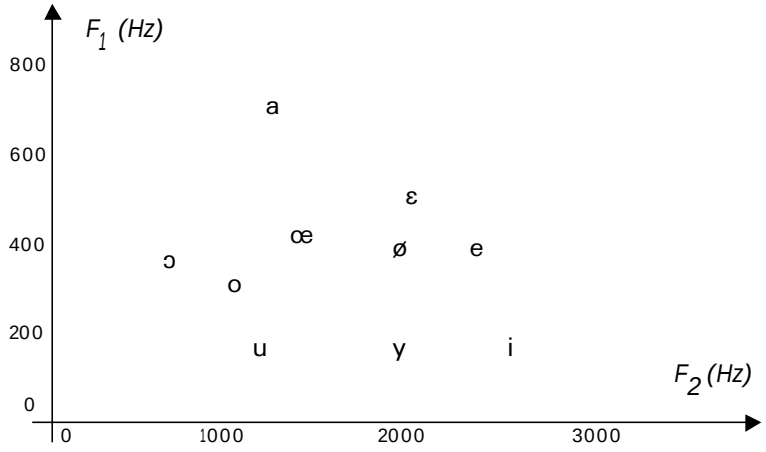


Fig. 1.13 Représentation des voyelles dans le plan F1-F2

1.2.2.4 Audition - perception

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production. En effet, tout ce qui peut être mesuré acoustiquement ou observé par la phonétique articulatoire n'est pas nécessairement perçu. Par ailleurs, nous avons déjà souligné, dans la section 1.1, le rôle fondamental que joue l'audition dans le processus même de production de la parole.

Les ondes sonores sont recueillies par l'appareil auditif, ce qui provoque les sensations auditives. Ces ondes de pression sont analysées dans l'***oreille interne*** qui envoie au cerveau l'influx nerveux qui en résulte; le phénomène physique induit ainsi un phénomène psychique grâce à un mécanisme physiologique complexe.

L'appareil auditif comprend l'***oreille externe***, l'***oreille moyenne***, et l'***oreille interne*** (Fig. 1.14). Le conduit auditif relie le pavillon au tympan : c'est un tube acoustique de section uniforme fermé à une extrémité, son premier mode de résonance est situé vers 3000 Hz, ce qui accroît la sensibilité du système auditif dans cette gamme de fréquences. Le mécanisme de l'oreille interne (***marteau***, ***étrier***, ***enclume***) permet une adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne. Les vibrations de l'étrier sont transmises au liquide de la ***cochlée***. Celle-ci contient la ***membrane basilaire*** qui transforme les vibrations mécaniques en impulsions nerveuses. La membrane s'élargit et s'épaissit au fur et à mesure que l'on se rapproche de l'apex de la cochlée; elle est le support de l'***organe de Corti*** qui est constitué par environ 25000 ***cellules ciliées*** raccordées au nerf auditif. La réponse en fréquence du conduit au droit de chaque cellule est esquissée à la figure 1.15. La fréquence de résonance dépend de la position occupée par la cellule sur la membrane; au-delà de cette

fréquence, la fonction de réponse s'atténue très vite. Les fibres nerveuses aboutissent à une région de l'écorce cérébrale appelée ***aire de projection auditive*** et située dans le lobe temporal. En cas de lésion de cette aire, on peut observer des troubles auditifs. Les fibres nerveuses auditives afférentes (de l'oreille au cerveau) et efférentes (du cerveau vers l'oreille) sont partiellement croisées : chaque moitié du cerveau est mise en relation avec les deux oreilles internes.

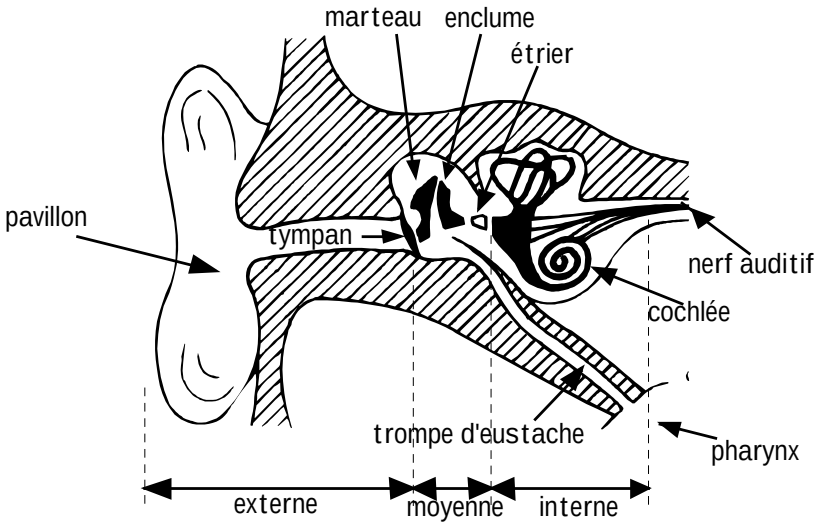


Fig. 1.14 Le système auditif.

Il reste très difficile de nos jours de dire comment l'information auditive est traitée par le cerveau. On a pu par contre étudier comment elle était finalement perçue, dans le cadre d'une science spécifique appelée ***psychoacoustique***. Sans vouloir entrer dans trop de détails sur la contribution majeure des psychoacousticiens dans l'étude de la parole, il est intéressant d'en connaître les résultats les plus marquants.

Ainsi, l'oreille ne répond pas également à toutes les fréquences. La figure 1.16 présente le champ auditif humain, délimité par la courbe de ***seuil de l'audition*** et celle du ***seuil de la douleur***. Sa limite supérieure en fréquence (≈ 16000 Hz, variable selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (≈ 32000 Hz).

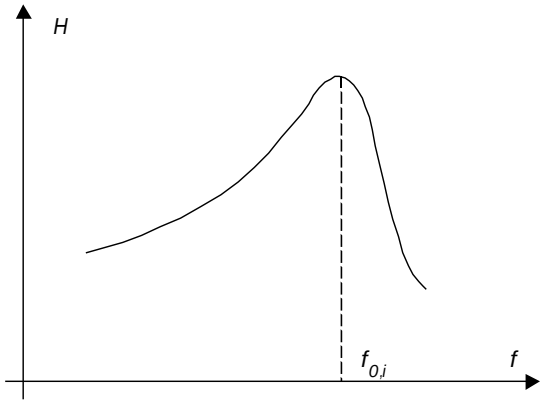


Fig. 1.15 Réponse en fréquence d'une cellule ciliée.

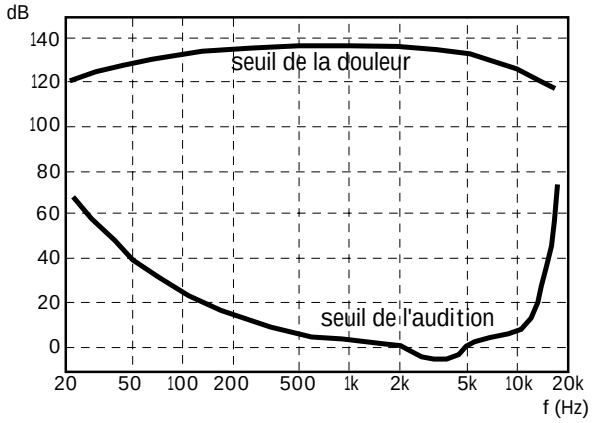


Fig. 1.16 Le champ auditif humain.

A l'intérieur de son domaine d'audition, l'oreille ne présente pas une sensibilité identique à toutes les fréquences. La figure 1.17.a fait apparaître les courbes d'égale impression de puissance auditive (aussi appelée ***sonie***, exprimée en ***sones***) en fonction de la fréquence. Elles révèlent un maximum de sensibilité dans la plage [500 Hz, 10 kHz], en dehors de laquelle les sons doivent être plus intenses pour être perçus.

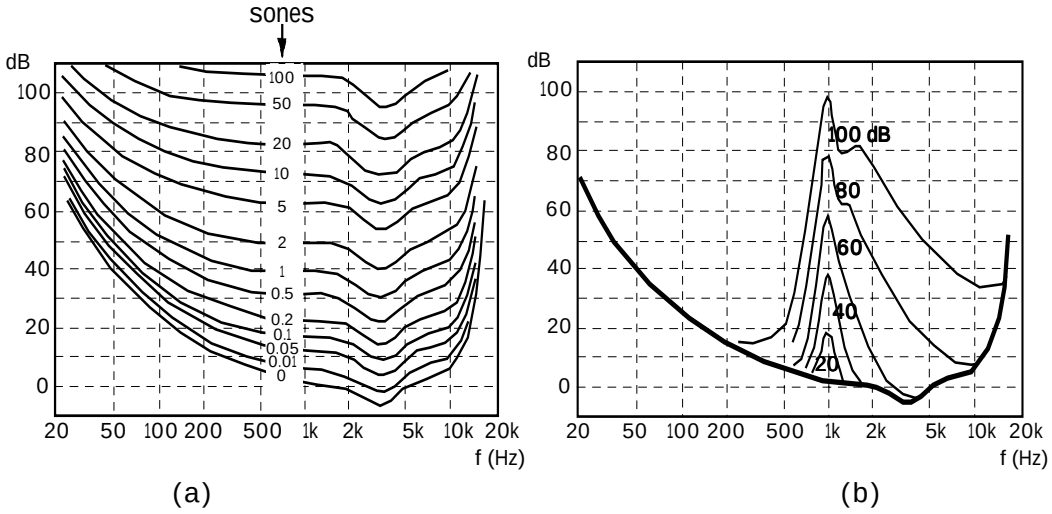


Fig. 1.17 (a) : Courbes isosoniques en champ ouvert. (b) : Masquage auditif par un bruit à bande étroite : limite d'audibilité en fonction de la puissance du bruit masquant.

Enfin, un son peut en cacher un autre. Cette propriété psychoacoustique, appelée ***phénomène de masquage***, peut être visualisée sous la forme de courbes de masquage (Fig. 1.17.b), qui mettent en évidence la modification locale du seuil d'audition en fonction de la présence d'un signal déterminé (un bruit à bande étroite centré sur 1 kHz dans le cas de la figure 1.17.b). Une modélisation efficiente des propriétés de masquage de l'oreille permet de réduire le débit binaire nécessaire au stockage ou à la transmission d'un signal acoustique, en éliminant les composantes inaudibles (voir chapitre 4).

Remarquons pour terminer que ce qui est perçu n'est pas nécessairement *compris.* Une connaissance de la langue interfère naturellement avec les propriétés psychoacoustiques de l'oreille. En effet, les sons ne sont jamais prononcés isolément, et le contexte phonétique dans lequel ils apparaissent est lui aussi mis à contribution par le cerveau pour la compréhension du message. Ainsi, certains sons portent plus d'information que d'autres, dans la mesure où leur probabilité d'apparition à un endroit donné de la chaîne parlée est plus faible, de sorte qu'ils réduisent l'espace de recherche pour les sons voisins. Les sons sont organisés en unités plus larges, comme les mots, qui obéissent eux-mêmes à une syntaxe et constituent une phrase porteuse de sens. Par conséquent, c'est tout notre savoir linguistique qui est mis à contribution lors du décodage acoustico-phonétique[8]. Les sections qui suivent ont précisément pour objet la description linguistique du signal de parole.

1.2.3 Le niveau phonologique

---

8 Cette influence marquée de la langue sur la perception fait elle aussi l'objet d'une étude spécifique, dans le cadre de la ***psycholinguistique***.

La ***phonologie*** (parfois appelée ***phonétique fonctionnelle***) est l'interface nécessaire entre la phonétique et les descriptions linguistiques de niveau plus élevé.

Dans les sections précédentes, nous avons décrit la parole comme si elle n'était porteuse d'aucune signification. Les sons de parole ont d'ailleurs été présentés indépendamment les uns des autres. La phonologie introduit la notion d'unité abstraite du discours (par opposition aux sons observés, perçus, ou articulés) : le ***phonème***. Le phonème est la plus petite unité phonique fonctionnelle, c'est-à-dire distinctive. Il n'est pas défini sur un plan acoustique, articulatoire, ou perceptuel, mais bien sur le plan fonctionnel. Ainsi, les phonèmes n'ont pas d'existence indépendante : ils constituent un ensemble structuré dans lequel chaque élément est intentionnellement différent de tous les autres, la différence étant à chaque fois porteuse de sens. La liste des phonèmes pour la plupart des langues européennes a été établie dès la fin du 19è siècle sur base de l'étude de ***paires minimales***, composées de paires de mots différant par un seul son, lequel suffit à changer leur sens (ex : [bɔ̃ - pɔ̃] dans '*bon-pont*').

La phonétique articulatoire pourrait donc être définie comme *l'étude de l'articulation de phonèmes*. Les phonèmes apparaissent en effet sous une multitude de formes articulatoires, appelées ***allophones*** (ou ***variantes***). Celles-ci résultent soit d'un changement volontaire dans l'articulation d'un son de base comme cela arrive souvent dans les prononciations régionales (ex : les différentes prononciations régionales du [ʀ] en français). De telles variations ne donnent pas naissance à de nouveaux phonèmes, puisqu'elles ne portent aucune information sémantique. Les variantes phoniques sont également causées, et ce de façon beaucoup plus systématique, par l'influence des phones environnants sur la dynamique du conduit vocal. Les mouvements articulatoires peuvent en effet être modifiés de façon à minimiser l'effort à produire pour les réaliser à partir d'une position articulatoire donnée, ou pour anticiper une position à venir. Ces effets sont connus sous le nom de ***réduction***, d'***assimilation***, et de ***coarticulation***. Les phénomènes coarticulatoires sont dus au fait que chaque articulateur évolue de façon continue entre les positions articulatoires. Ils apparaissent même dans le parlé le plus soigné (Fig. 1.18). Au contraire, la réduction et l'assimilation prennent leur origine dans des contraintes physiologiques et sont sensibles au débit parlé. L'assimilation est causée par le recouvrement de mouvements articulatoires et peut aller jusqu'à modifier un des traits phonétiques du phonème prononcé[9]. La réduction est plutôt due au fait que les cibles articulatoires sont moins bien atteintes dans le parler rapide. Certains phonèmes, comme les semi-voyelles, les liquides, et le plosives, y sont plus sensibles. Les phénomènes de *réduction*, d'*assimilation*, et de *coarticulation* sont

---

[9] Deux cas bien connus d'assimilation en français sont l'assimilation de nasalité et l'assimilation de sonorité. On les produit d'ailleurs précisément en les citant : *'assimilation de sonorité'* se prononce [asimilasjõrsonorite] au lieu de [asimilasjõdsonorite], ce qui constitue précisément une assimilation de sonorité. De même pour *'assimilation de nasalité',* prononcé [asimilasiõnnazalite] au lieu de [asimilasjõdnazalite], qui constitue lui-même une assimilation de nasalité.

en grande partie responsables de la complexité des traitements réalisés sur les signaux de parole pour en obtenir l'analyse, la reconnaissance, ou la synthèse.

Les représentations phonémiques font également usage de l'IPA, mais les symboles sont entourés de séparateurs obliques ('/') plutôt que de crochets. La distinction entre phonologie et phonétique apparaît clairement lorsqu'on aligne des séquences phonémiques avec leurs expressions phonétiques : les variations phoniques n'apparaissent que dans les dernières. Ainsi, en français, /ʀ/ peut être voisé [ʀ] ou pas [ʀ̥] (*parent* - /paʀã/ - [paʀã], mais *pitre* - /pitrʀ/ - [pitʀ̥]). En anglais, /ŋ/ est parfois prononcé [n], surtout en fin de mot (*something* - /sʌmθɪŋ/ - [sʌmθɪn]). Il est important de bien comprendre que les symboles phonétiques ont pour rôle de transcrire ce qui a été dit, que ce soit porteur de sens ou pas. Si une nuance peut être perçue, il doit y avoir un moyen de la transcrire phonétiquement, mais elle n'apparaîtra pas nécessairement dans une transcription phonémique. D'ailleurs, les phonéticiens disposent d'un nombre important de marques diacritiques à ajouter aux symboles de l'API pour rendre compte des variations de lieu et de mode articulatoire (comme le marqueur de dévoisement [̥], de voisement [̬] ou de prononciation dentale [̪]).

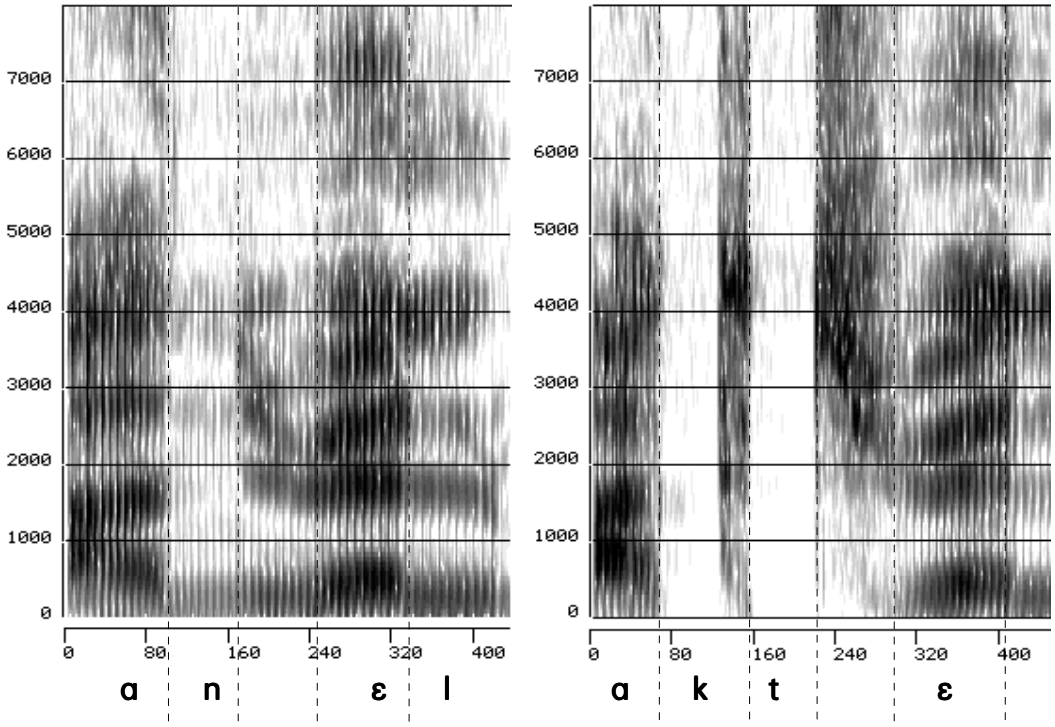


Fig. 1.18 Un cas d'assimilation de sonorité (coarticulation affectant le voisement d'une sonore). A gauche, le début du mot *'annuellement'*, dans lequel [ɥ] est placé dans un contexte voisé. A droite, le début de '*actuellement*' : [ɥ] est totalement dévoisé à cause de la plosive sourde qui précède.

Notons pour terminer qu'une description phonologique ne peut être complète si elle ne permet pas de rendre compte de la durée, de l'intensité, et de la

fréquence fondamentale des phonèmes, dans la mesure où ces grandeurs apportent au message parlé une information qui ne se retrouve pas dans les symboles de l'API. Ces trois composantes sont collectivement désignées sous le terme de ***prosodie***. La durée des silences et des phones détermine le ***rythme*** de la phrase, tandis que l'évolution de la fréquence fondamentale constitue sa ***mélodie***. Cependant, la définition d'unités prosodiques abstraites (que l'on pourrait appeler ***prosodèmes***) soulève de nombreuses questions, qui restent aujourd'hui encore sans réponse définitive. Il n'existe pas à ce jour d'*alphabet prosodique international*, ni de méthode de transcription prosodique universellement admise. Tout au plus dispose-t-on dans l'API d'un petit nombre de signes. La durée est indiquée par des marques d'allongement (allongement [ ] ou semi-allongement [·]). L'API ne fournit pas de moyen efficient pour la transcription des mouvements d'intensité ou de fréquence fondamentale, si ce n'est pas sous la forme de marqueurs d'***accent*** (primaire [ˈ] ou secondaire [ˌ]), défini comme la mise en évidence d'une voyelle obtenue en la prononçant plus fort, plus longue, ou non-réduite. Nous reparlerons de ce problème au chapitre 7, consacré à la synthèse de la parole.

1.2.4 Le niveau morphologique

La suite des phonèmes prononcés correspond à des mots, choisis dans le ***lexique*** des mots de la langue. Si l'on sait que le Petit Robert compte à peu près 50.000 entrées et que seules les formes canoniques y sont répertoriées (masc. sing. des noms et adjectifs, infinitifs des verbes), on peut estimer la richesse lexicale d'une langue comme le français à plusieurs centaines de milliers de mots.

Lorsqu'on étudie les formes écrites et phonétiques d'une langue, il est frappant de constater que les mots qui la composent, bien que très nombreux, sont eux-mêmes constitués d'unité plus petites (comme dans *image, images, imagine, imagination, imagerie, inimaginable,* etc.*)*. La ***morphologie*** est la branche de la linguistique qui étudie comment les formes lexicales sont obtenues à partir d'un ensemble réduit d'unités porteuses de sens, appelées ***morphèmes***. On distingue les morphèmes lexicaux des morphèmes grammaticaux, qui apportent aux premiers des nuances de genre, nombre, mode, temps, personne, etc. Tout comme le phonème, le morphème est une unité abstraite. Elle peut être réalisée en pratique sous diverses formes appelées ***allomorphes***, fonction de leur contexte morphémique. Ainsi le morphème grammatical du pluriel se manifeste-t-il sous la forme d'un '*s*' dans '*pommes*', d'un '*x*' dans '*jeux*' et d'un '*nt*' dans '*jouent*'.

On fait généralement la différence entre la morphologie ***inflectionnelle***, qui compte des caractéristiques morphologiques telles que le genre, le nombre, le mode, le temps, la personne, etc. (*image, images)*, la morphologie ***dérivationnelle***, qui étudie la construction des mots de catégories syntaxiques différentes à partir d'un morphème de base (*image, imagine, imagination, imagerie*) et la morphologie ***compositionnelle***, dont le rôle est d'expliquer comment plusieurs morphèmes peuvent se composer pour en former un nouveau (*in+imaginable=inimaginable*).

L'importance de la morphologie en traitement de la parole tient à ce que la catégorie grammaticale et la prononciation des mots peuvent être expliquée dans une large mesure par leur composition morphémique.

1.2.5 1.2.5. Le niveau syntaxique

Toute suite de mots du lexique ne forme pas une phrase correcte. En effet, la liste des phrases admises, bien qu'infinie dans les langues naturelles, est restreinte par leur ***syntaxe***. Ceci constitue d'ailleurs la définition du mot *syntaxe*¸ qu'il ne faut pas confondre avec les règles utilisées pour la décrire, organisées sous la forme de ***grammaires***. Les mots du lexique y perdent leur individualité pour n'être plus vus qu'en tant que ***parties du discours*** (ou *natures*), listes de mots interchangeables pour une grammaire donnée). Par exemple, la grammaire (arbitrairement simplifiée) :

phrase = groupe nominal + verbe conjugué

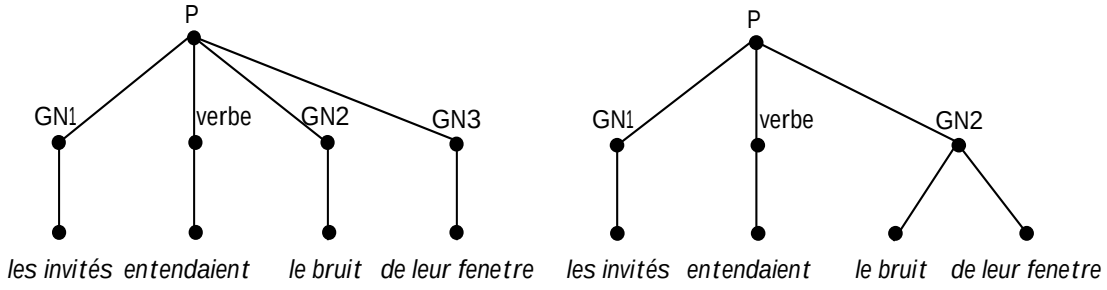
groupe nominal = déterminant + nom [+ préposition + groupe nominal]

où les crochets indiquent des composantes optionnelles, interdit les phrases *déterminant+verbe conjugué* comme dans *'mon donne'* ou *'les joue'*.

Les grammaires ne servent pas qu'à dresser la frontière entre les phrases régulièrement constituées ou pas : elles permettent également de décrire l'organisation hiérarchique des phrases : leur ***structure syntaxique***. L'information qui en résulte possède un avantage appréciable sur la représentation linéaire des phrases : elle met en évidence leur(s) structure(s) interne(s) possible(s). Par exemple, la suite de mots :

'Les invités entendaient le bruit de leur fenêtre'

pourra être mise en correspondance avec les structures syntaxiques suivantes :



Notons pour terminer qu'il est en général possible de rendre compte de la syntaxe d'un langage avec plusieurs grammaires, en fonction du type des règles utilisées. Ainsi, la grammaire Grévisse du français n'est qu'une description syntaxique parmi beaucoup d'autres. Elle se prête par ailleurs très mal à une implantation logicielle, car elle suppose une connaissance et un usage préalable de la langue. Le verbe y est par exemple défini comme '*le mot qui exprime, soit l'action faite ou subie par le sujet, soit l'union de l'attribut au sujet*'. A l'inverse, les grammaires dites ***formelles*** sont apparues dès les années 50 (dans le cadre

d'une science nouvelle, appelée ***linguistique informatique*** ou ***computationnelle***) dans le but de formaliser la syntaxe des langues naturelles sous une forme strictement utilisable par l'ordinateur. Cet objectif n'a jamais été complètement atteint[10].

1.2.6 Le niveau sémantique

Si la syntaxe restreint l'ensemble de phrases acceptables pour une langue donnée, elle ne constitue cependant pas une limite exhaustive d'acceptabilité. En effet, bon nombre de phrases syntaxiquement correctes restent inadmissibles (ex : *'la politesse jaune pleure du pain'*). Cette imprécision tient à la confusion qui est faite, par les grammaires, des mots appartenants a une même liste d'éléments du discours.

L'étude des significations des mots, de la façon dont elles sont liées les unes aux autres, et des bases du choix lexical fait l'objet de la ***sémantique lexicale***. Parmi les principales questions qu'il lui appartient d'examiner, les problèmes d'ambiguïté de portée prennent un part importante. Une phrase aussi simple que :

'Jean-François n'est pas parti à New York en avion'.

peut en effet être comprise comme :

Quelqu'un d'autre est parti à New York en avion

Jean-François est parti de New York en avion

Jean-François est parti ailleurs.

Jean-François est parti à New York par un autre moyen de transport

selon l'étendue du champ d'application de la négation, et ceci bien que toutes ces acceptions admettent la même description syntaxique.

Lorsqu'on attribue aux mots du lexique des ***traits sémantiques*** (abstrait/concret, animé/inanimé, couleur, forme, etc.), le nombre de classes de mots interchangeables augmente rapidement (et le nombre de mots par classe tombe en proportion), de sorte que pratiquement chaque mot retrouve son individualité. Les règles qui décrivent l'organisation d'une phrase en fonction des relations entre les traits sémantiques des mots qui la constituent deviennent alors nettement plus complexes qu'au niveau purement syntaxique.

---

10 Il reste que les grammaires formelles sont utilisées de façon intensive dans les langages informatiques. Elles ont également été augmentées de caractéristiques sémantiques qui en font de puissants outils pour le traitement du langage naturel. Enfin, les grammaires formelles et leurs extensions stochastiques ont reçu un intérêt considérable ces vingt dernières années, principalement dans le cadre du traitement automatique des structures linguistiques. Nous les aborderons plus en profondeur au chapitre 6.

Notons que la différence entre sémantique et syntaxe reste relativement floue[11]. Ainsi, une description syntaxique est souvent porteuse de sens (voir l'exemple de *"les invités entendaient le bruit de leur fenêtre"*). D'une façon générale, toute analyse syntaxique basée sur un nombre important de classes d'éléments du discours possède inévitablement un caractère sémantique.

L'étude de grammaires et d'analyseurs sémantiques est un sujet de recherche actuel en linguistique informatique. Seules des solutions partielles ont pu être obtenues jusqu'à présent (analyse dans un domaine sémantique restreint).

1.2.7 Le niveau pragmatique (ou niveau du discours)

Au contraire du sens sémantique, que l'on qualifie souvent d'*indépendant du contexte*, le sens ***pragmatique*** est défini comme *dépendant du contexte.* Tout ce qui se réfère au contexte, souvent implicite, dans lequel une phrase s'inscrit et à la relation entre le locuteur et de son auditoire, a quelque chose à voir avec la pragmatique[12]. Son étendue couvre l'étude de sujets tels que les *présuppositions*, les *implications de dialogue*, les *actes de parole indirects*, etc. Elle est malheureusement bien moins développée encore que la sémantique.

---

11 *Stricto sensu*, cette distinction apparaît plus clairement lorsque la sémantique est définie comme l'étude des conditions de vérité de propositions logiques. Ce n'est pas notre propos ici.

12 Dans un large mesure, la pragmatique est utilisée pour "*balayer tous les aspects complexes liés à la signification, dont les chercheurs veulent remettre l'examen à plus tard*". Ceci contribue sans aucun doute à la difficulté que l'on éprouve lorsqu'on veut établir la frontière entre sémantique et pragmatique.

CHAPITRE 2
MODELISATION LPC ET CODAGE DE LA PAROLE

2.1 Information - Redondance - Variabilité

Le signal vocal est caractérisé par une très grande redondance, condition nécessaire pour résister aux perturbations du milieu ambiant. Pour aborder la notion de redondance, il faut examiner la parole en tant que vecteur d'information.

On peut établir grossièrement une classification de l'information vocale en trois catégories : le *sens* du message délivré, tel qu'obtenu par l'analyse sémantique, les *nuances* qui y sont apportées par le locuteur, essentiellement grâce à la prosodie, et le *locuteur* lui-même, dont la marque apparaît tant dans le timbre de la voix (liée à l'évolution temporelle de l'enveloppe spectrale) que dans la prosodie utilisée[13]. L'anlalyse sémantique est quant à tributaire à la fois de l'évolution de l'enveloppe spectrale et de la prosodie. En effet, la reconnaissance des phonèmes est essentiellement basée sur un examen spectral, alors que l'organisation des phonèmes en mots du lexique, et de mots en groupes syntaxiques ou sémantiques, met à profit la relation étroite qui unit prosodie et syntaxe (ou sémantique). A cet égard, la prosodie est souvent considérée comme un indice acoustique qui doit permettre à l'auditeur de percevoir rapidement la structure de la phrase.

En admettant que l'on ne s'intéresse qu'au sens de la phrase, et que l'on néglige le rôle joué par la prosodie dans sa compréhension, on peut réduire l'information

[13]La caractéristique principale d'une imitation réussie est la modification conjointe du timbre et de l'intonation de l'imitateur. Souvent, d'ailleurs, la seule copie de l'intonation suffit à faire reconnaître la personne imitée.

qu'elle apporte à la suite des phonèmes qui la composent. Un calcul simple permet alors d'estimer le débit binaire correspondant.

Considérons un message constitué d'une suite de caractères x_i d'un alphabet $X=[x_1, x_2, \ldots x_L]$. Si $p(x_i)$ est la probabilité associée à l'occurrence a priori de x_i, son apparition apporte une information :

$$I = -\log_2 p(x_i)$$

De sorte que l'information apportée en moyenne par l'apparition d'un caractère quelconque de X est donnée par :

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

C'est l'*entropie* de la source qui délivre les suites de caractères x_i. Pour la langue française, dont la probabilité d'occurrence des phonèmes est connue, on obtient H=4.73 (si les 37 phonèmes étaient équiprobables, on aurait trouvé H= 5.21, puisque $2^{5.21} = 37$).

Par conséquent, si l'on admet que, dans la conversation courante, environ 10 phonèmes sont prononcés par seconde, l'information moyenne est inférieure à 50 bits/s. Ce chiffre est à comparer au débit binaire maximum admissible sur un canal téléphonique, qui fait aujourd'hui partie des connaissances de tout un chacun : la plupart des modems envoient des informations numériques sur nos lignes téléphoniques avec un débit binaire de 33.6 kbits/s. On en conclut que le signal vocal est extrêmement redondant.

Cette redondance sera mise à profit par les techniques de *codage de la parole*, dont le but sera de diminuer le débit nécessaire au stockage ou à la transmission de la parole sans nuire à son intelligibilité.

En pratique, cependant, il ne s'agit pas de redondance stricte, puisque nous avons négligé les nuances apportées par la prosodie et les caractéristiques propres à chaque locuteur. On parlera plutôt dans ce cas de *variabilité.* Le problème des techniques de *reconnaissance de la parole* sera précisément de retrouver le sens d'une phrase malgré l'extrême variabilité qui la caractérise.

2.2 Modélisation

L'analyse de la parole est une étape indispensable à toute application de synthèse, de codage, ou de reconnaissance. Elle repose en général sur un *modèle*. Celui-ci possède un ensemble de *paramètres* numériques, dont les plages de variation définissent l'ensemble des signaux couverts par le modèle. Pour un signal et un modèle donné, l'*analyse* consiste en l'*estimation* des paramètres du modèle dans le but de lui faire correspondre le signal analysé. Pour ce faire, on met en oeuvre un *algorithme d'analyse*, qui cherche généralement à minimiser la différence, appelée *erreur de modélisation*, entre le signal original et celui qui serait produit par le modèle s'il était utilisé en tant que synthétiseur (fig. 2.1).

Les erreurs de modélisation peuvent être causées par le modèle lui-même, puisqu'il restreint implicitement l'ensemble des signaux couverts, de sorte qu'un signal qui n'en fait pas partie ne pourra jamais être modélisé correctement (quelque soit la valeur choisie pour les paramètres). On parlera alors d'*erreur intrinsèque*. D'autre part, un signal couvert par le modèle ne sera effectivement bien modélisé que dans la mesure où l'algorithme d'analyse est effectivement capable de trouver les paramètres qui annulent l'erreur de modélisation. Si ce n'est pas le cas, on parlera d'*erreur extrinsèque*, ou *biais d'analyse*.

Il existe de nombreux modèles de parole. On distingue les modèles *articulatoires*, les modèles *de production*, et les modèles *phénoménologiques* :

- Les premiers réalisent une simulation numérique du mécanisme de phonation. Leurs paramètres sont essentiellement de nature articulatoire (position de la langue, ouverture des lèvres,...). La parole est décrite comme le résultat du passage d'un flux d'air à travers un ensemble de tubes de section variable. L'analyse par modèle articulatoire, également appelée *inversion acoustique articulatoire*, est un problème complexe. Les algorithmes font intervenir les équations de la mécanique des fluides.
- A la différence des précédents, les modèles de production ne cherchent à reproduire que le schéma de principe du mécanisme phonatoire, par le biais de son équivalent électrique. On y décrit la parole comme le signal produit par un assemblage de générateurs et de filtres numériques. Les paramètres de ces modèles sont ceux des générateurs et filtres qui les constituent. Le modèle Auto-Régressif (AR), que nous étudierons plus spécialement dans ce chapitre en raison de utilisation répandue en traitement de la parole, en est l'exemple le plus simple.
- Enfin, les modèles phénoménologiques cherchent à modéliser le signal de parole sans se soucier de la façon dont il a été produit. Les algorithmes d'analyse qui y sont associés se rapportent par conséquent plus au traitement du signal en général qu'au traitement de la parole. Les modèles basés sur l'analyse de Fourier en sont un exemple.

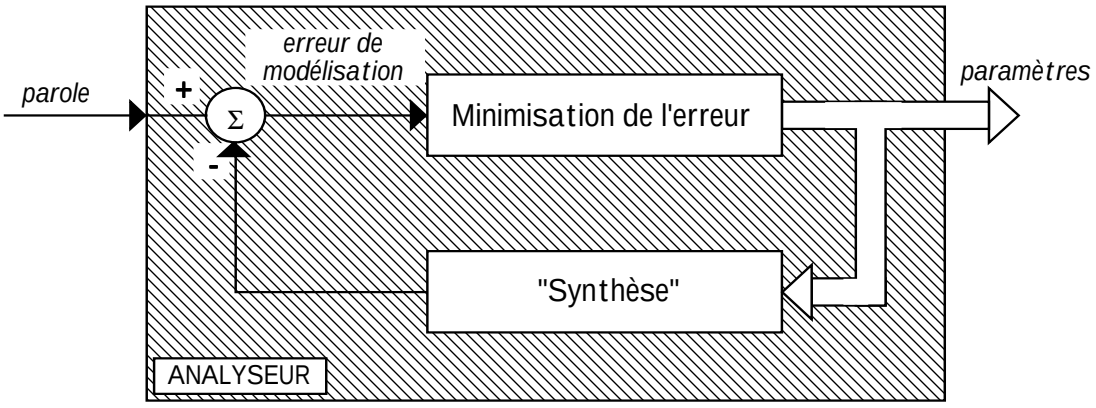


Fig. 2.1 Schéma de principe d'un analyseur de parole. En pratique, l'étape de synthèse peut être implicite.

2.3 Un modèle électrique de la phonation : le modèle AutoRégressif (AR)

Fant a proposé en 1960 un modèle de production dont nous résumons ici la version numérique.

Un signal voisé peut être modélisé par le passage d'un train d'impulsions *u(n)* à travers un filtre numérique récursif de type *tout pôles*. On montre que cette modélisation reste valable dans le cas de sons non-voisés, à condition que *u(n)* soit cette fois un bruit blanc. Le modèle final est illustré à la figure 2.3. Il est souvent appelé *modèle auto-régressif*, parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme :

$$x(n) = \sigma u(n) + \sum_{i=1}^{p} -a_i x(n-i) \qquad (2.5)$$

(où *u(n)* est le signal d'excitation), ce qui exprime que chaque échantillon est obtenu en ajoutant un terme d'excitation à une prédiction obtenue par combinaison linéaire de *p* échantillons précédents. Les coefficients du filtre sont d'ailleurs appelés *coefficients de prédiction* et le modèle AR est souvent appelé *modèle de prédiction linéaire*.

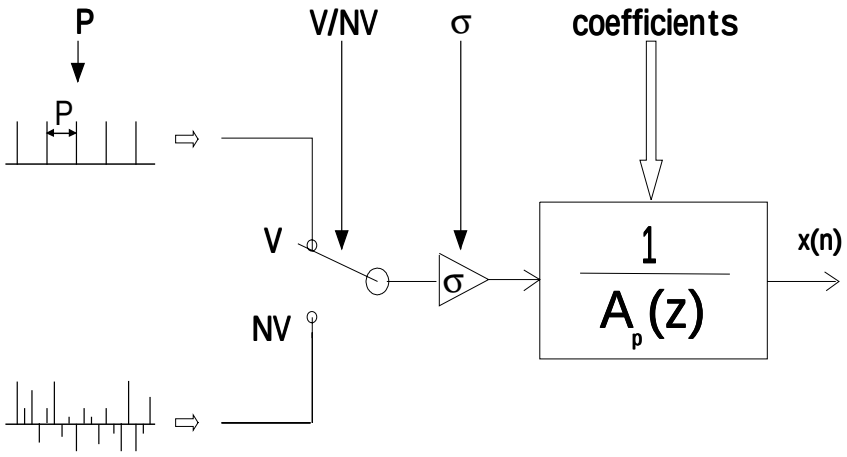


Fig. 2.3 Le modèle auto-régressif.

Les paramètres[14] du modèle AR sont : la période du train d'impulsions (sons voisés uniquement), la décision Voisé/NonVoisé (V/NV), le gain σ, et les coefficients du filtre *1/A(z)*, appelé *filtre de synthèse*.

---

[14] On trouvera sur le site web du cours (http://tcts.fpms.ac.be/cours/1005-08/speech/) un didacticiel (LPCLearn) permettant de mieux comprendre le rôle de chacun des paramètres de ce modèle.

Le problème de l'estimation d'un modèle AR, souvent appelée *analyse LPC* (pour 'Linear Prediction Coding'[15]) revient à déterminer les coefficients d'un filtre tout pôles dont on connaît le signal de sortie, mais pas l'entrée. Il est par conséquent nécessaire d'adopter un critère, afin de faire un choix parmi l'infinité de solutions possibles. Le critère classiquement utilisé est celui de la *minimisation de l'énergie de l'erreur de prédiction*. La mise en équation de ce problème conduit aux équations dites de *Yule-Walker* :

$$\Phi \mathbf{a} = -\phi \tag{2.13}$$

avec

$$\phi = \left[\phi_{xx}(1), \phi_{xx}(2), \cdots, \phi_{xx}(p)\right]^T$$

$$\mathbf{a} = \left[a_1, a_2, \cdots, a_p\right]^T$$

$$\Phi = \begin{Bmatrix} \phi_{xx}(0) & \phi_{xx}(1) & \phi_{xx}(2) & \cdots & \phi_{xx}(p-1) \\ \phi_{xx}(1) & \phi_{xx}(0) & \phi_{xx}(1) & \cdots & \phi_{xx}(p-2) \\ \phi_{xx}(2) & \phi_{xx}(1) & \phi_{xx}(0) & \cdots & \phi_{xx}(p-3) \\ \vdots & \vdots & \vdots & & \vdots \\ \phi_{xx}(p-1) & \phi_{xx}(p-2) & \phi_{xx}(p-3) & \cdots & \phi_{xx}(0) \end{Bmatrix} \tag{2.14}$$

On constate en passant que la matrice Φ est symétrique et que les diagonales parallèles à la diagonale principale contiennent des éléments égaux. Une telle matrice est dite *Toeplitz.* Il existe dans ce cas des algorithmes rapides pour la résolution, appelés algorithmes de Levinson et de Schur.

Considérations pratiques

Pour mener à bien une analyse LPC, il faut pouvoir choisir :

- la fréquence d'échantillonnage f_e;
- la méthode d'analyse et l'algorithme correspondant;
- l'ordre p de l'analyse LPC;
- le nombre d'échantillons par tranche N et le décalage entre tranche successives L;

Le choix de la fréquence d'échantillonnage est fonction de l'application visée et de la qualité du signal à analyser. On choisira plutôt 8 kHz pour les signaux téléphoniques, 10 kHz pour les applications de reconnaissance, et 16 kHz pour les applications de synthèse. Dans le cadre d'applications multimédia, on préférera les fréquences normalisées de 11.25 et 22.5 kHz, sous-multiples des 44.1 kHz du Compact Disk.

L'ordre d'analyse conditionne le nombre de formants que l'analyse est capable de prendre en compte. On estime en général que la parole présente un formant par kHz de bande passante, ce qui correspond à une paire de pôles pour $A_p(z)$. Si on y ajoute une paire de pôles pour la modélisation de l'excitation glottique, on

[15] La prédiction linéaire a été initialement utilisée pour le codage de la parole.

obtient les valeurs classiques de *p=10, 12, et 18* pour f_e*=8, 10 et 16 kHz* respectivement. Elles trouvent d'ailleurs une justification expérimentale dans le fait que l'énergie de l'erreur de prédiction diminue rapidement lorsqu'on augmente *p* à partir de 1, pour tendre vers une asymptote autour de ces valeurs : il devient inutile d'encore augmenter l'ordre, puisqu'on ne prédit rien de plus.

La durée des tranches d'analyse et leur décalage sont souvent fixées à 30 et 10 ms respectivement. Ces valeurs ont été choisies empiriquement; elles sont liées au caractère quasi-stationnaire du signal de parole.

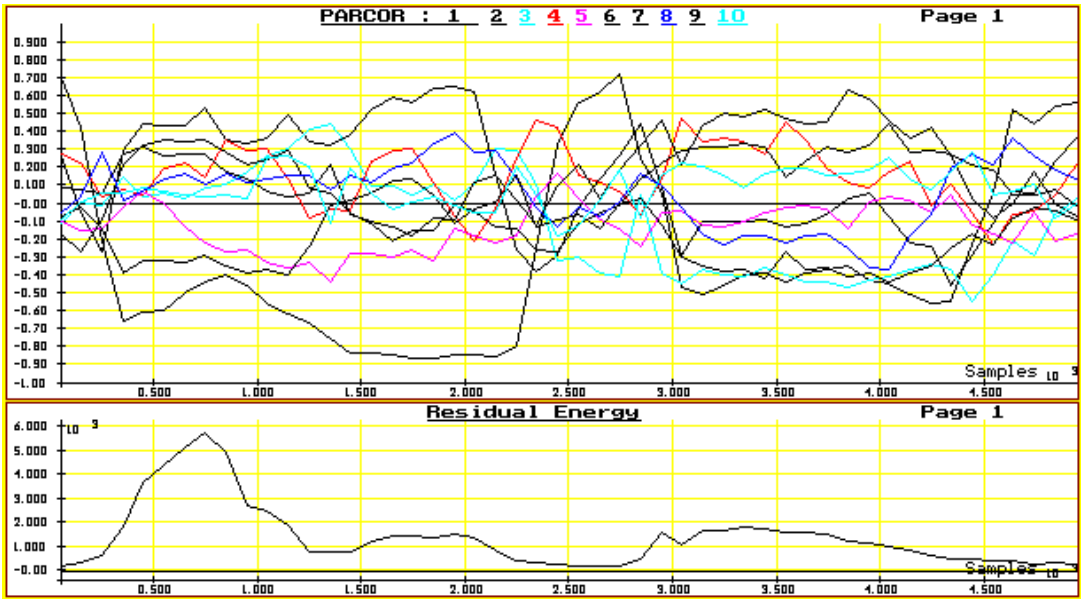
Enfin, pour compenser les effets de bord, on multiplie en général préalablement chaque tranche d'analyse par une fenêtre de pondération *w(n)* de type *fenêtre de Hamming* :

$$w(n) = 0.54 + 0.46\cos(2\pi\frac{n}{N}) \quad \text{pour} \quad n=0\ldots N-1 \qquad (2.44)$$

On retiendra donc que l'analyse LPC d'un signal de parole implique la résolution d'un système de (l'ordre de) **10 d'équations à 10 inconnues toutes les 10 ms.**

2.4 Un exemple complet

La figure 2.15. donne la représentation AR du mot *'parenthèse'* (F_e=8kHz, *p*=10), telle qu'obtenue après analyse (cfr. 2.2) par prédiction linéaire (LPC : Linear Prediction Coding). L'analyse est menée sur des tranches de 30 ms (240 échantillons), à raison d'une analyse toutes les 10 ms.



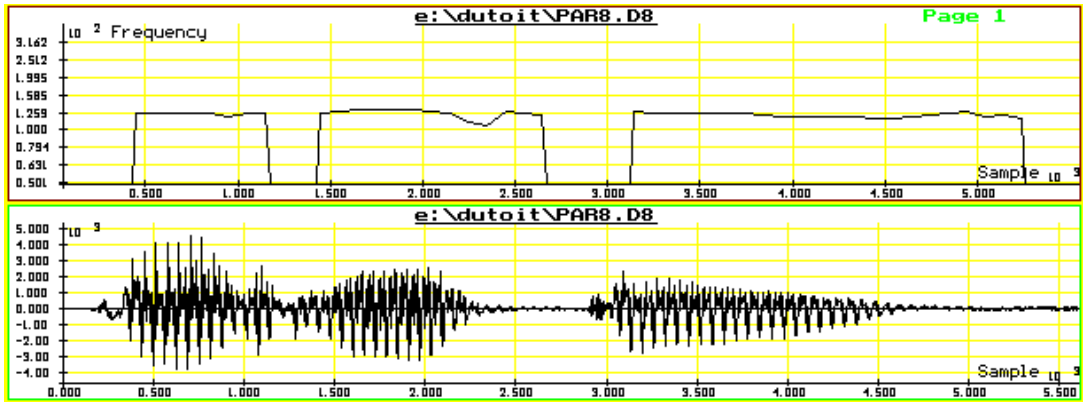


Fig. 2.15 Analyse LPC du mot *'parenthèse'*. De haut en bas : les coefficients PARCOR, le gain σ, et le pitch *P*. La décision V/NV apparaît implicitement dans la représentation du pitch.

2.5 Codage LPC

La figure ci-dessous donne le schéma de principe d'un codeur LPC, tel qu'il peut être utilisé pour les transmissions de voix par satellite (ex : voix d'un journaliste en mission dans un pays lointain) ou plus communément dans un GSM. La signal vocal mesuré par le micro est découpé en trames, analysé par l'algorithme de Schur et par un algorithme d'analyse de la fréquence des cordes vocales. Les paramètres qui en résultent sont *quantifiés,* c.-à-d. qu'ils sont codés sur un ensemble fini de nombres entiers (ce qui permet d'associer à chaque paramètre un nombre fini de *bits* par trame).

En d'autres termes, lors d'un appel par GSM, le GSM émetteur (qui n'est rien d'autre qu'un ordinateur de poche spécialisé dans l'analyse, le codage, le décodage, et la synthèse LPC) enregistre la parole à transmettre, en réalise toutes les 10 ms une analyse LPC (par laquelle il trouve les coefficients de prédiction qui « collent » le mieux au conduit vocal de l'appelant, pour la tranche de parole considérée), et transmet ces coefficients (et non la voix originale de l'appelant). Le GSM récepteur reçoit quant à lui les paramètres du conduit vocal de l'appelant, produit un signal de synthèse simulant ce conduit vocal, et le fait entendre au correspondant, qui croit entendre l'appelant. Il s'agit pourtant bien de parole de synthèse, au même titre qu'on pourrait imaginer une caméra inspectant l'appelant et ne transmettant d'un modèle 3D de don visage, lequel serait reproduit en image de synthèse coté récepteur.

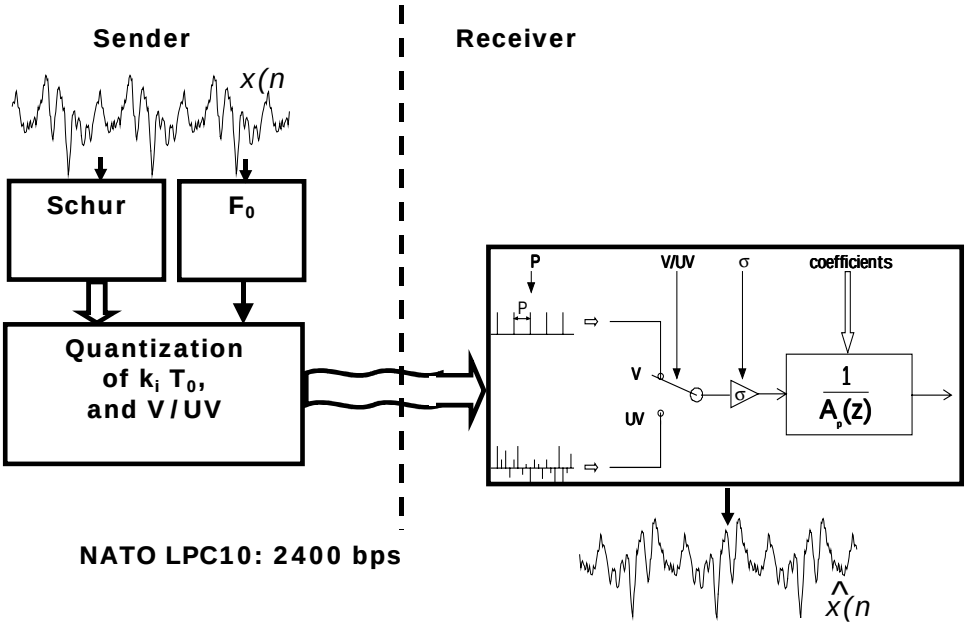


Fig. 2.15 Transmission de parole basée sur le codage LPC

CHAPITRE 3
SYNTHESE DE LA PAROLE

3.1 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, que ce texte ait été directement introduit par un opérateur sur un clavier alpha-numérique, qu'il ait été scanné et reconnu par un système de reconnaissance optique des caractères (OCR : *Optical Character Recognition*), ou qu'il ait été produit automatiquement par un système de dialogue homme-machine. Un tel système diffère fondamentalement d'autres machines parlantes en ceci qu'il est destiné à donner lecture de 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 des annonceurs vocaux automatiques dans les gares. On définira donc plutôt la synthèse TTS comme la *production automatique de phrases par calcul de leur transcription phonétique.*

3.2 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 a 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. Le célèbre astrophysicien Stephen Hawking donne tous ses cours à l'université de Cambridge de cette façon. La synthèse offre également des services aux personnes mal-voyantes, en leur donnant accès à l'information écrite "en noir"[16], à condition de coupler le synthétiseur à un logiciel de reconnaissance des caractères.

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).

Monitoring vocal. Dans certains cas, l'information orale est plus efficace qu'un message écrit. L'utilisation d'une voix de synthèse dans un centre de contrôle de site industriel, par exemple, permet d'attirer l'attention du personnel de surveillance sur un problème urgent. De la même manière, l'intégration d'un synthétiseur dans la cabine de pilotage d'un avion permet d'éviter au pilote d'être dépassé par la quantité d'informations visuelles qu'il a à analyser. Et quand on voit à quoi ressemblera bientôt le tableau de bord de nos voitures, on comprend qu'elles ne tarderont pas à nous parler. La maison de demain, quant à elle, fera bien de se doter d'une voix de synthèse si elle veut avertir ses occupants d'une anomalie constatée sur un de ses circuits de surveillance.

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

---

16 Dans le vocabulaire des aveugles, l'impression "en noir" s'oppose à l'impression en Braille.

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.

3.3 Analyse du problème

Le profane considère souvent la synthèse de la parole comme un problème assez trivial. C'est une tâche que nous effectuons tous sans le moindre effort apparent. De là à dire que, étant donné l'état actuel des connaissances et des techniques, et vu les progrès récemment acquis en traitement du signal et en traitement du langage naturel, il doit être possible à un ordinateur d'égaler l'homme dans ce domaine, il n'y a qu'un pas... que nous nous garderons bien de franchir ici. Le processus de lecture puise en effet au plus profond des ressources, souvent insoupçonnées, de l'intelligence humaine. Il suffit pour s'en convaincre de constater qu'il est rare qu'un enfant lise de façon naturelle avant l'âge de six ans, c'est à dire lorsqu'il a acquis presque toutes ses autres fonctions intellectuelles. L'usage de la parole est d'ailleurs lui-même assez tardif; il est toujours précédé de sa reconnaissance et de sa compréhension.

Qu'il soit donc clair dès à présent que la suite des opérations à effectuer pour obtenir la lecture automatique d'une phrase ne se conformera que de loin au schéma fonctionnel tout naturellement adopté par le cerveau. 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. Comme le fait très justement remarquer Lindblom: "*Après tout, les avions ne battent pas des ailes !*"

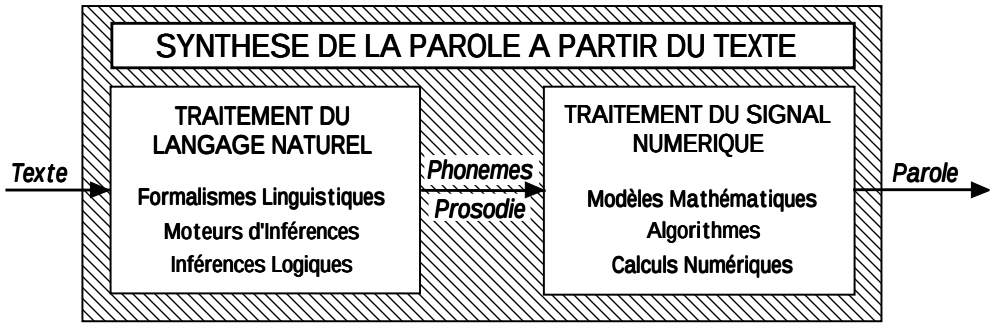


Fig. 7.1 Diagramme fonctionnel d'un système de synthèse TTS.

La figure 7.1 donne le diagramme fonctionnel d'un synthétiseur TTS. On y retrouve un bloc de traitement du langage naturel, capable de produire la transcription phonétique de la phrase à lire et d'y associer une intonation et un rythme naturels, et un module de traitement du signal, homologue de l'appareil phonatoire, qui transforme cette information symbolique en signal de parole.

3.4 Organisation générale du module de traitement du langage naturel

L'organisation générale des opérations de traitement du langage réalisées par le synthétiseur est donnée à la figure 7.2. On y remarque immédiatement, outre la présence attendue des modules de ***phonétisation automatique*** et de ***génération de la prosodie***, l'importance du module d'***analyse morpho-syntaxique***. De fait, l'extraction d'informations morphologiques, et par là de la nature des mots et de leur organisation syntaxique, est indispensable pour au moins deux raisons :

1. On ne peut réaliser de transcription phonétique correcte en ignorant la nature des mots à prononcer ou le lien de dépendance syntaxique entre mots successifs.
2. La prosodie partage un lien privilégié avec la syntaxe. Il est clair qu'elle dépend également du sens de la phrase et de son contexte (informations sémantiques et pragmatiques), mais vu le peu de données disponibles sur ces relations, les systèmes TTS se concentrent essentiellement sur la syntaxe. Le plus souvent, même, de façon assez superficielle.

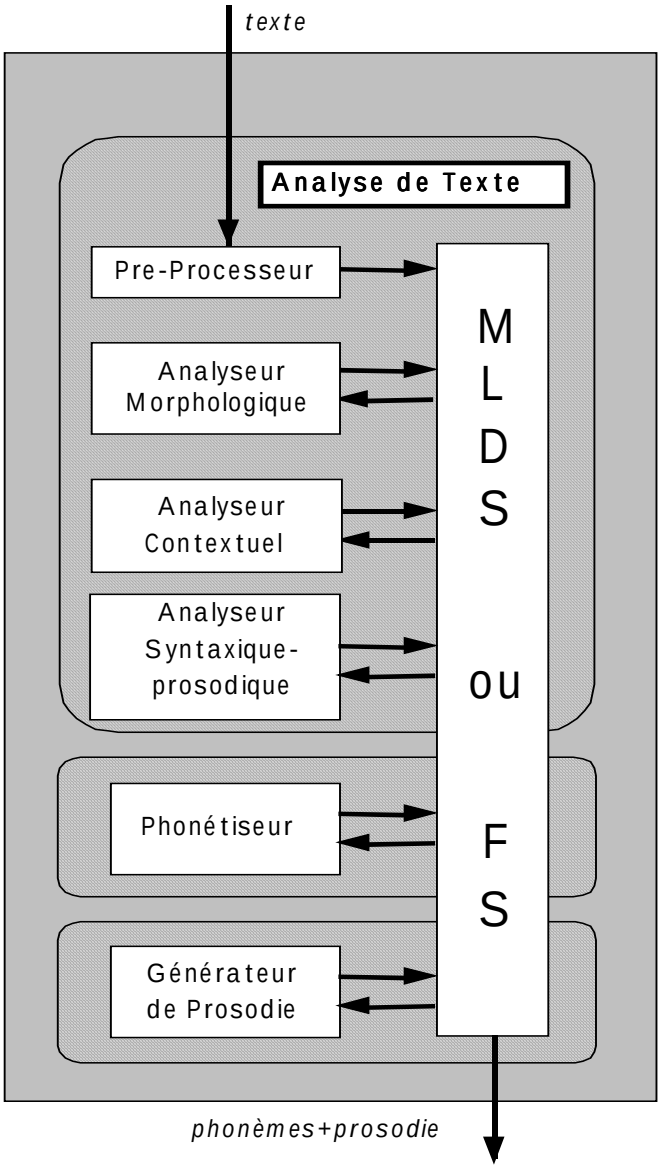


Fig. 7.2 Le module de traitement du langage naturel d'un système de conversion texte-parole.

Le module d'analyse morpho-syntaxique de la figure 7.2 est lui-même composé de :

- Un module de ***prétraitement***, qui joue principalement le rôle d'interface entre le texte (représentation linéaire) et la structure de donnés internes gérée par le synthétiseur. Ce module identifie toutes les séquences de caractères qui risquent de poser un problème de prononciation : nombres, abréviations, acronymes, expressions toutes faites, etc. et les transcrit éventuellement en toutes lettres.

- Un ***analyseur morphologique***, qui a pour tâche de proposer toutes les natures possibles pour chaque mot pris individuellement, en fonction de sa graphie.
- Un ***analyseur contextuel***, qui considère les mots dans leur contexte, ce qui lui permet de réduire la liste des natures possibles pour chaque mot en fonction des natures possibles des mots voisins.
- Enfin, un ***analyseur syntaxique-prosodique***, qui examine l'espace de recherche restant et établit un découpage du texte en groupes de mots qui permettra d'y associer une prosodie.

Il est cependant de plus en plus courant d'organiser les données internes de façon structurée. Lorsqu'on cherche à classer les systèmes TTS les plus récents, deux grandes tendances apparaissent : ceux qui mettent en œuvre les ***structures d'attributs*** (*feature structures* : FS) présentées au Chapitre 6, en associant à chaque niveau linguistique une catégorie PATR (comme par exemple SVOX), et ceux qui utilisent des ***structures de données multi-niveaux*** (*multi-level data structure* : MLDS, Fig. 7.3), où chaque niveau apparaît comme une description de la phrase indépendante mais synchronisée avec les autres (comme dans le système SPEECH MAKER, dans EULER et dans FESTIVAL, sous l'impulsion du système DELTA). Une telle représentation structurée de l'information accroît considérablement la ***lisibilité*** des données (et donc des modules de traitement qui y accèdent) et leur ***extensibilité***. Les structures d'attributs et les structures multi-niveaux admettent en effet toutes deux la ***sous-spécification*** : il est toujours possible d'ajouter des catégories ou des niveaux sans interférer avec ceux qui existent déjà.

Les problèmes posés par l'analyse morphosyntaxique, la transcription phonétique automatique, et la génération de la prosodie ne seront pas traités ici. Nous abordons par contre dans la section suivante le principe de base des techniques de traitement du signal utilisées pour la mise en forme du signal de synthèse.

3.5 Synthèse par règles - Synthèse par concaténation

On peut établir une analogie fonctionnelle entre le rôle joué par le module de traitement du signal de la figure 7.1 et celui du système phonatoire humain, qui contrôle en permanence l'activité de tous ses muscles (y compris de ceux qui règlent la fréquence de vibration des cordes vocales) de façon à produire le signal voulu. Pour y arriver, il est clair que ce module doit, dans une certaine mesure, prendre en compte les contraintes articulatoires[17]. On sait depuis longtemps en effet que les transitions phonétiques contribuent plus à l'intelligibilité du signal vocal que les zones stables des phonèmes. On peut alors envisager de le faire de deux façons :

---

17 Et ce, même si les techniques de synthèse les plus récentes décrivent la parole sous la forme d'une séquence temporelle de paramètres qui n'ont pas de lien direct avec les traits articulatoires.

- De façon explicite, sous la forme d'une série de règles décrivant formellement l'influence des phones les uns sur les autres;
- De façon implicite, en enregistrant des exemples de transitions entre phones dans une base de données de segments de parole, et en les utilisant tels quels comme unités de parole (en lieu et place des phones).

Cette alternative a donné lieu à deux grandes familles de synthétiseurs, voire à deux *philosophies* de synthèse, vu leurs divergences de moyens et d'objectifs : la ***synthèse par règles*** et *la* ***synthèse par concaténation***.

3.5.1 Synthèse par règles

Les ***synthétiseurs par règles*** ont principalement la faveur des phonéticiens et des phonologistes. Ils permettent une approche cognitive, générative du mécanisme de la phonation. 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.

Les synthétiseurs par règles sont organisés comme à la figure 7.12.

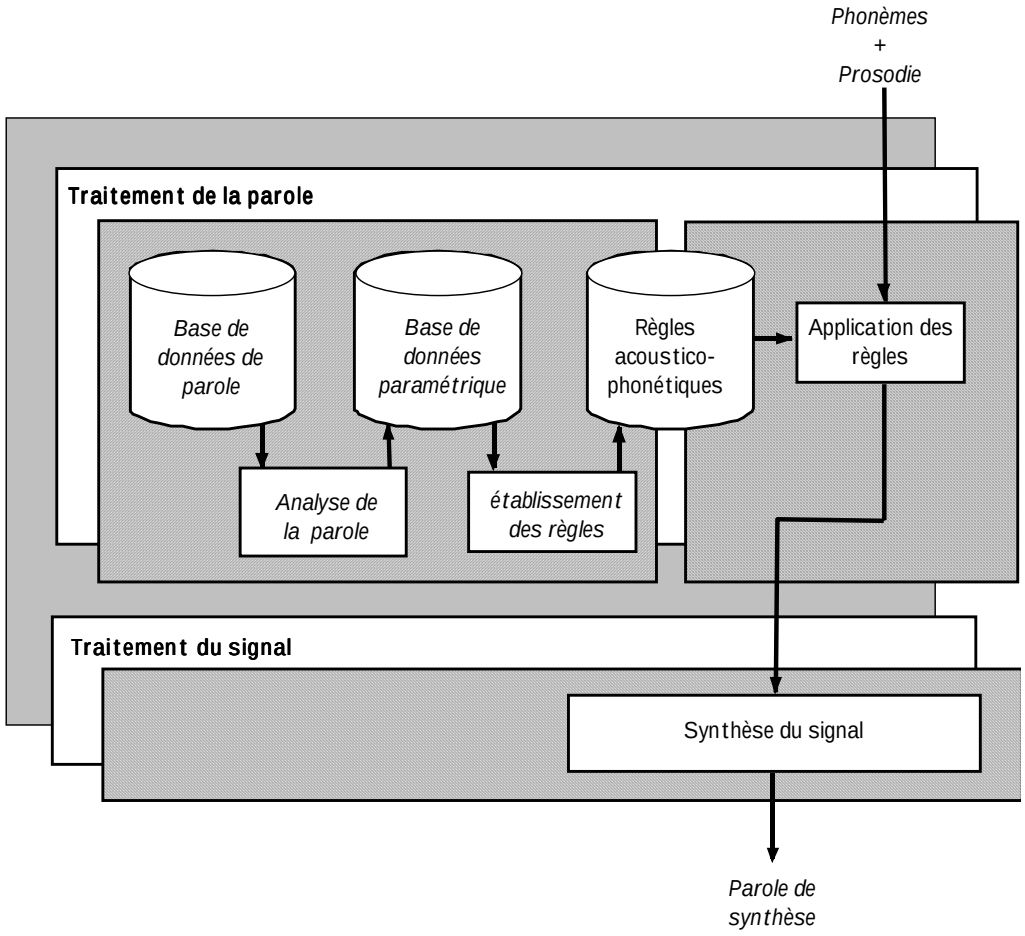


Figure 7.12 Schéma de conception et fonctionnement typique d'un système de synthèse par règles. La plupart des traitements se situent effectivement dans le domaine du traitement de la parole, puisqu'ils font intervenir des connaissances qui lui sont spécifiques. Le bloc en haut à gauche correspond à la mise au point du synthétiseur, les autres se rapportent à sa mise en œuvre.

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. 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, plus propice à l'établissement des règles. Celles-ci sont généralement proposées par des phonéticiens. On commence par inspecter globalement l'ensemble des données, de façon à établir la forme générale des règles à produire. 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. Il est

à remarquer que cette étape d'estimation est menée sur une seule voix : un moyennage inter-locuteur aurait peu de signification dans ce contexte. De même, les règles provenant de synthétiseurs déjà existants ne peuvent resservir que dans la mesure où elles modélisent des caractéristiques articulatoires générales plutôt que des particularités du locuteur ayant enregistré le corpus (sauf bien entendu si l'on cherche a produire des règles caractérisant précisément le passage d'une voix à une autre). La mise au point du synthétiseur s'achève par une long processus d'essais-erreurs, afin d'optimiser la qualité de la synthèse.

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 une 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.

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. La qualité des voix disponibles n'est en effet pas aussi bonne qu'en synthèse pas concaténation, pour un coût de développement supérieur.

3.5.2 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 paroles mises en œuvre par le synthétiseur. Ceci apparaît clairement dans la description générale d'un tel synthétiseur sur la figure 7.17, 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 son* ne faisant aucune référence explicite à la nature profonde des signaux traités, au contraire du bloc de *traitement de la parole* qui suppose des connaissances phonétiques : ces opérations pourraient en effet tout aussi bien être mises en œuvre, par exemple, pour la synthèse de signaux musicaux. 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* (7.6.2) qui s'accompagne d'un *lissage* purement *acoustique* des discontinuités au droit des 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.

On commence ainsi par faire choix d'un ensemble d'unités de parole qui devront permettre de minimiser les futurs problèmes de concaténation. 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)[18], 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 enferment assez correctement les phénomènes de coarticulation tout en ne nécessitant qu'un nombre limité d'unités.

18 Dans le cas de phonèmes ne présentant pas de partie stationnaire, soit on prend la partie le plus stable, soit on fait appel à un triphone, ce qui évite de devoir segmenter dans du transitoire.

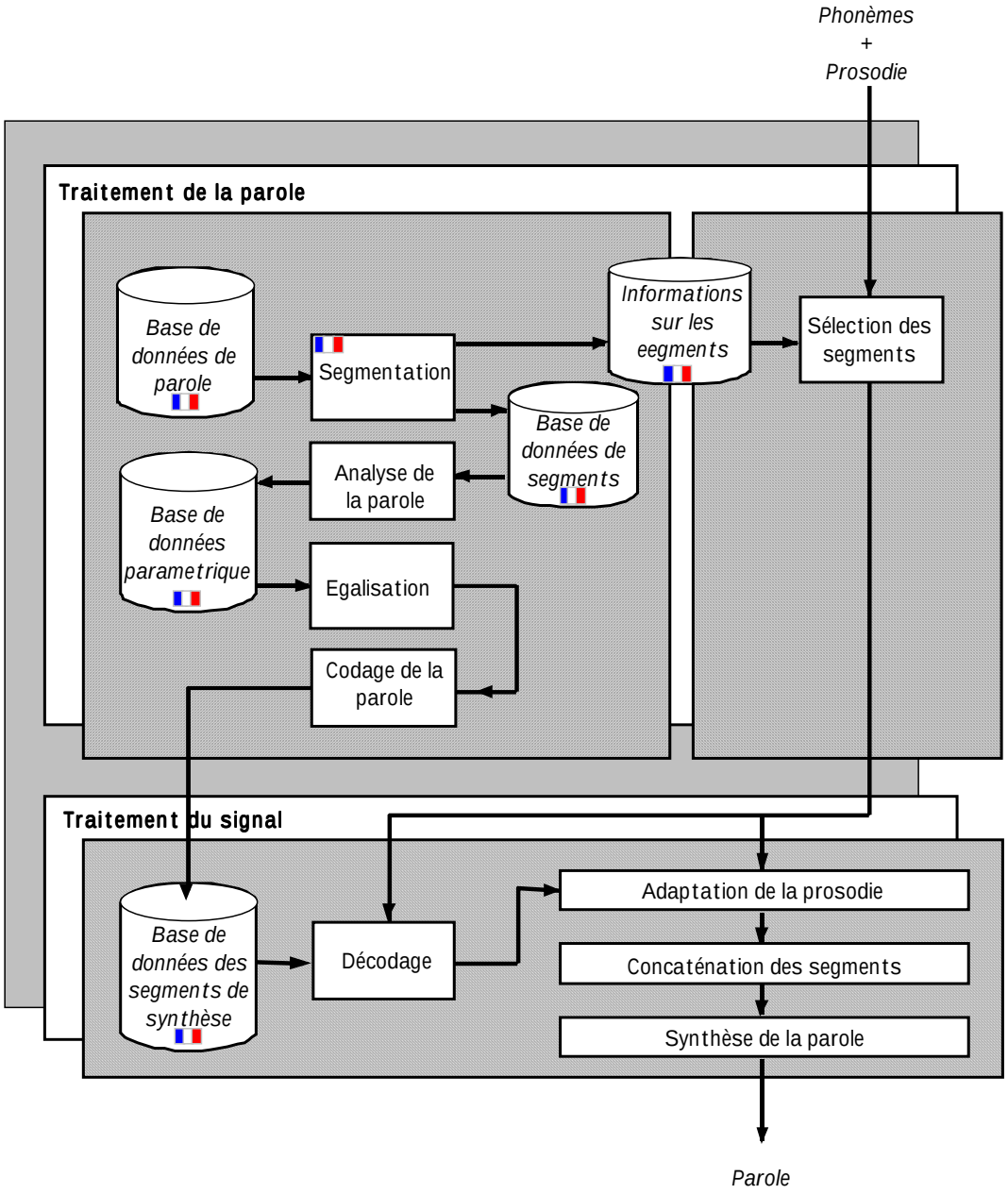


Fig. 7.17 Schéma général d'un synthétiseur par concaténation. Les opérations qui dépendent de la langue sont indiquées par un drapeau.

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 a 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* 7.6.2). 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 dont les décisions sont ensuite vérifiées et éventuellement corrigées manuellement. Le résultat de cette segmentation constitue la ***base de données de segments***[19], qui comprend les échantillons de tous les segments utilisables. On centralise également l'information relative à ces segments (leur nom, leur durée, leur pitch, et les marqueurs de frontières de phonèmes à l'intérieur des segments; dans 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) dans une base de données séparée, qui sera utilisée par le bloc de sélection d'unités.

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*** (ex : LPC) 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 à la figure 7.17. 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

---

19 Souvent, on mémorise les segments avec entre 50 et 100 ms de contexte gauche et droit, de façon à éviter l'apparition de transitoires lors de leur analyse paramétrique.

deux problèmes fondamentaux évoqués plus haut: la modification de la prosodie des unités et leur concaténation.

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 7.17. La distinction introduite plus haute entre sélection *segmentale* d'unités et sélection *totale* (segmentale et supra-segmentale) 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. 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[20]: 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***. 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[21]*** apparus récemment, qui procèdent par sélection totale et dynamique. Au moment de choisir les segments à mettre en œuvre, 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 discontinuité spectrale l'une par rapport à l'autre. 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

---

[20] Ce seul cas couvre la grande majorité des systèmes de synthèse par concaténation actuellement commercialisés.

[21] Ces unités sont appelées de la sorte en raison du fait que, si l'on règle l'algorithme de sélection de façon à ce qu'il favorise le choix d'unités consésécutives dans la base de données (ce qui rend inutile une concaténation des ces unités), tout se passe comme si on procédait par concaténation d'unités de taille variable (diphones, triphones, syllabes, morceaux de mots, voire mots entiers).

segments choisi correspondent-ils au contexte phonétique et prosodique dans lequel on les insère?) et d'un ***coût de concaténation*** (dans quelle mesure la juxtaposition des segments choisis amène-t-elle des discontinuités).

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. Puisque les segments sont en général 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[22]).

3.5.2.1 Exemple : Synthèse LPC

Le modèle LPC se prête particulièrement bien aux étapes de modification de la prosodie et de concaténation, requises pour la mise en forme d'un signal respectant la commande phonétique d'entrée et exempt de « clicks » de concaténation.

La modification de la durée est tout simplement réalisée en synthétisant plus ou moins d'échantillons avec les mêmes coefficients de prédiction. On obtient par exemple un signal deux fois plus long que l'original en produisant 20 ms de signal pour chaque décalage de 10 ms sur le signal original.

La modification de l'intonation est tout aussi triviale, puisque F0 est un paramètre explicite du modèle : il suffit de changer ce paramètre à la valeur imposée en entrée pour produire un signal ayant la fréquence requise.

La concaténation peut être produite en « lissant » les paramètres du filtre de part et d'autre du point de concaténation (voir fig. ci-dessous).

---

22 Ce qui explique d'ailleurs la complexité des systèmes de synthèse par règles; en synthèse par concaténation, les discontinuités spectrales sont suffisamment faibles pour que le lissage réalisé sur les paramètres de modèles acoustiques paraisse naturel (mais suffisamment importantes pour nécessiter un lissage malgré tout).

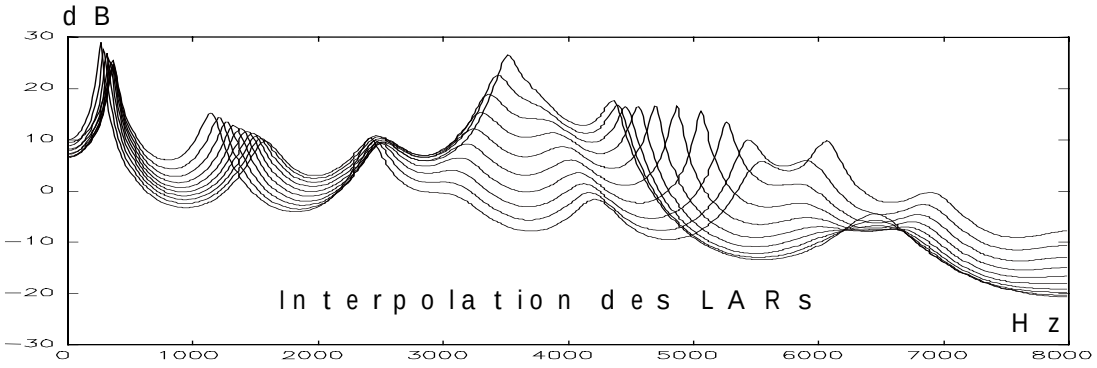


Fig. 7.21 Lissage linéaire des *log area ratios* entre les phonèmes /l/ des diphones /al/ et /lo/ (p=18).

3.5.2.2 Exemple : Synthèse dans le domaine temporel

On a vu apparaître ces dernières années des méthodes basées sur une modification temporelle directe de la forme du signal. L'idée sous-jacente est qu'il est possible de modifier l'intonation et la durée d'un signal sans l'usage d'aucun modèle paramétrique, en évitant ainsi toute possibilité d'erreur de modélisation.

Si *s(n)* est un signal purement périodique, il est en effet possible d'en obtenir un signal $\tilde{s}(n)$ de même enveloppe spectrale que *s(n)* mais de fréquence fondamentale différente en additionnant des ***fenêtres d'OLA*** $s_i(n)$, extraites par multiplication de $s(n)$ par une fenêtre de pondération *w(n)* synchronisée sur le pitch T_0 de *s(n)*. La modification de fréquence fondamentale se fait en changeant l'écartement temporel entre fenêtres d'OLA successives (de sa valeur T_0 de départ à une valeur T quelconque), et en réadditionnant les unes aux autres les fenêtres d'OLA ainsi écartées (Fig. 7.26):

$$s_i(n) = s(n)\, w(n - i\, T_0) \tag{7.42}$$

$$\tilde{s}(n) = \sum_{i=-\infty}^{\infty} s_i(n - i(T - T_0)) \tag{7.43}$$

L'équation (7.43) fournit un moyen très simple de modifier la fréquence fondamentale d'un signal périodique (Fig. 7.27).

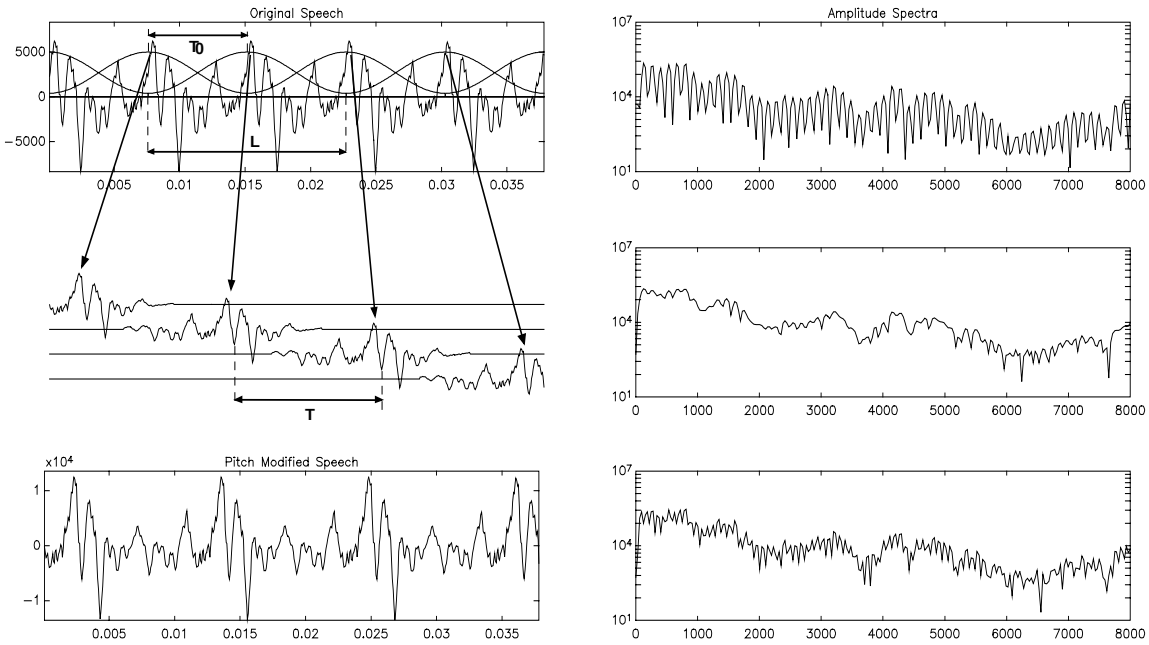


Fig. 7.27 Le processus de ré-harmonisation spectrale de TD-PSOLA. A gauche, les signaux, à droite les spectres correspondants. Le signal modifié (en bas) a bien la même enveloppe spectrale que le signal de départ (en haut); mais pas la même fréquence fondamentale.

Les méthodes de synthèse utilisant ce principe mathématique, comme TD-PSOLA, MBROLA, LP-PSOLA sont maintenant très utilisées en synthèse par concaténation. Elles requièrent en effet une très faible charge de calcul (temps réel sur Intel386) et leur qualité segmentale est excellente.

CHAPITRE 4
RECONNAISSANCE DE LA PAROLE

4.1 Introduction

Le problème de la reconnaissance automatique de la parole consiste à extraire, à l'aide d'un ordinateur, l'information lexicale contenue dans un signal de parole. Depuis plus de deux décennies, des recherches intensives dans ce domaine ont été accomplies par de nombreux laboratoires internationaux. Des progrès importants ont été accomplis grâce au développement d'algorithmes puissants ainsi qu'aux avancées en traitement du signal.

Différents systèmes de reconnaissance de la parole ont été développés, couvrant des domaines aussi vastes que la reconnaissance de quelques mots clés sur lignes téléphoniques, les systèmes à dicter vocaux, les systèmes de commande et contrôle sur PC, et allant jusqu'aux systèmes de compréhension du langage naturel (pour applications limitées).

Dans ce chapitre, nous discutons des principes qui sont à la base de la plupart de ces systèmes.

Malgré que nous ayons appris beaucoup concernant la reconnaissance de la parole et la mise en oeuvre de systèmes pratiques et utiles, il reste encore beaucoup de questions fondamentales concernant la technologie pour lesquelles nous n'avons pas de réponses. Il est clair que le signal de parole est un des signaux les plus complexes. En plus de la complexité physiologique inhérente au système phonatoire et des problèmes de coarticulation qui en résultent, le conduit vocal varie également très fort d'une personne à l'autre.

Enfin, la mesure de ce signal de parole est fortement influencée par la fonction de transfert (comprenant les appareils d'acquisition et de transmission, ainsi que l'influence du milieux ambiant).

4.2 Niveaux de complexité

Pour bien appréhender le problème de la reconnaissance automatique de la parole, il est bon d'en comprendre les différents niveaux de complexités et les différents facteurs qui en font un problème difficile.

Il y a d'abord le problème de la *variabilité intra et inter-locuteurs*. Le système est-il *dépendant du locuteur* (optimisé pour un locuteur bien particulier) ou *indépendant du locuteur* (pouvant reconnaître n'importe quel utilisateur)?

Evidemment, les systèmes dépendants du locuteur sont plus faciles à développer et sont caractérisés par de meilleurs taux de reconnaissance que les systèmes indépendants du locuteur étant donné que la variabilité du signal de parole est plus limitée. Cette dépendance au locuteur est cependant acquise au prix d'un entraînement spécifique à chaque utilisateur. Ceci n'est cependant pas toujours possible. Par exemple, dans le cas d'applications téléphoniques, il est évident que les systèmes doivent pouvoir être utilisés par n'importe qui et doivent donc être indépendants du locuteur. Bien que la méthodologie de base reste la même, cette indépendance au locuteur est cependant obtenue par l'acquisition de nombreux locuteurs (couvrant si possible les différents dialectes) qui sont utilisés simultanément pour l'entraînement de modèles susceptibles d'en extraire toutes les caractéristiques majeures. Une solution intermédiaire parfois utilisée est de développer des systèmes capable de s'adapter (de façon supervisée ou non supervisée) rapidement au nouveau locuteur.

Le système reconnaît-il des *mots isolés* ou de la *parole continue*? Evidemment, il est plus simple de reconnaître des mots isolés bien séparés par des périodes de silence que de reconnaître la séquence de mots constituant une phrase. En effet, dans ce dernier cas, non seulement la frontière entre mots n'est plus connue mais, de plus, les mots deviennent fortement articulés (c'est-à-dire que la prononciation de chaque mot est affectée par le mot qui précède ainsi que par celui qui suit - un exemple simple et bien connu étant les liaisons du français). Dans le cas de la parole continue, le niveau de complexité varie également selon qu'il s'agisse de texte lu, de texte parlé ou, beaucoup plus difficile, de langage naturel avec ses hésitations, phrases grammaticalement incorrectes, faux départs, etc. Un autre problème, qui commence à être bien maîtrisé, concerne la *reconnaissance de mots clés* en parole libre. Dans ce dernier cas, le vocabulaire à reconnaître est relativement petit et bien défini mais le locuteur n'est pas contraint de parler en mots isolés. Par exemple, si un utilisateur est invité à répondre par « oui » ou « non », il peut répondre « oui, s'il vous plait ». Dans ce contexte, un problème qui reste particulièrement difficile est le rejet de phrases ne contenant aucun mots clés.

La *taille du vocabulaire* et son *degré de confusion* sont également des facteurs importants. Les petits vocabulaires sont évidemment plus faciles à reconnaître que les grands vocabulaires, étant donné que dans ce dernier cas, les possibilités de confusion augmentent. Certains petits vocabulaires peuvent cependant s'avérer particulièrement difficiles à traiter; ceci est le cas, par exemple, pour l'ensemble des lettres de l'alphabet, contenant surtout des mots très courts et acoustiquement proches.

Le système est-il *robuste,* c.-à-d. capable de fonctionner proprement dans des *conditions difficiles*? En effet, de nombreuses variables pouvant affecter significativement les performances des systèmes de reconnaissance ont été identifiées:

- Les bruits d'environnement tels que bruits additifs stationnaires ou non-stationnaires (par exemple, dans une voiture ou dans une usine).
- Acoustique déformée et bruits (additifs) corrélés avec le signal de parole utile (par exemple, distorsions non linéaires et réverbérations).
- Utilisation de différents microphones et différentes caractéristiques (fonctions de transfert) du système d'acquisition du signal (filtres), conduisant généralement à du bruit de convolution.
- Bande passante fréquentielle limitée (par exemple dans le cas des lignes téléphoniques pour lesquelles les fréquences transmises sont naturellement limitées entre environ 350Hz et 3200Hz).
- Elocution inhabituelle ou altérée, comprenant entre autre: l'effet Lombard, (qui désigne toutes les modifications, souvent inaudibles, du signal acoustique lors de l'élocution en milieu bruité), le stress physique ou émotionnel, une vitesse d'élocution inhabituelle, ainsi que les bruits de lèvres ou de respiration.

Certains systèmes peuvent être plus robustes que d'autres à l'une ou l'autre de ces perturbations, mais en règle générale, les reconnaisseurs de parole actuels restent encore trop sensibles à ces paramètres.

4.3 Principes généraux

Le problème de la reconnaissance automatique de la parole consiste à extraire l'information contenue dans un signal de parole (signal électrique obtenu à la sortie d'un microphone et typiquement échantillonné à 8kHz dans le cas de lignes téléphoniques ou entre 10 et 16kHz dans le cas de saisie par microphone). Bien que ceci soulève également le problème de la *compréhension de la parole*, nous nous contenterons ici de discuter du problème de la reconnaissance des mots contenus dans une phrases.

4.3.1 Reconnaissance par comparaison à des exemples

Les premiers succès en reconnaissance vocale ont été obtenus dans les années 70 à l'aide d'un paradigme de reconnaissance de mots « par l'exemple ». L'idée, très simple dans son principe, consiste à faire prononcer un ou plusieurs exemples de chacun des mots susceptibles d'être reconnus, et à les enregistrer sous forme de *vecteurs acoustiques* (typiquement : un vecteur de coefficients LPC ou assimilés toutes les 10 ms). Puisque cette suite de vecteurs acoustiques caractérisent complètement l'évolution de l'enveloppe spectrale du signal enregistré, on peut dire qu'elle correspond à un l'enregistrement d'un spectrogramme. L'étape de reconnaissance proprement dite consiste alors à analyser le signal inconnu sous la forme d'une suite de vecteurs acoustiques similaires, et à comparer la suite inconnue à chacune des suites des exemples

préalablement enregistrés. Le mot « reconnu » sera alors celui dont la suite de vecteurs acoustique (le « spectrogramme ») colle le mieux à celle du mot inconnu. Il s'agit en quelque sorte de voir dans quelle mesure les spectrogrammes se superposent.

Ce principe de base n'est cependant pas implémentable directement : un même mot peut en effet être prononcé d'une infinité de façons différentes, en changeant le rythme de l'élocution. Il en résulte des spectrogramme plus ou moins distordus dans le temps. La superposition du spectrogramme inconnu aux spectrogramme de base doit dès lors se faire en acceptant une certaine « élasticité » sur les spectrogrammes candidats. Cette notion d'élasticité est formalisée mathématiquement par un algorithme désormais bien connu : l'algorithme DTW (*Dynamic Time Warping*, en anglais).

ON comprend aisément qu'une telle technique soit intrinsèquement limitée par la taille du vocabulaire à reconnaître (une centaine de mots tout au plus) et qu'elle soit plus propice à la reconnaisance monolocuteur (une reconnaissance multilocuteur imposerait d'enregistrer, de stocker, et surtout d'utiliser pour la comparaison, de nombreux exemples pour chaque mot). Les résultats obtenus, dans le contexte monolocuteur/petit vocabulaire, sont aujourd'hui excellents (proches de 100%).

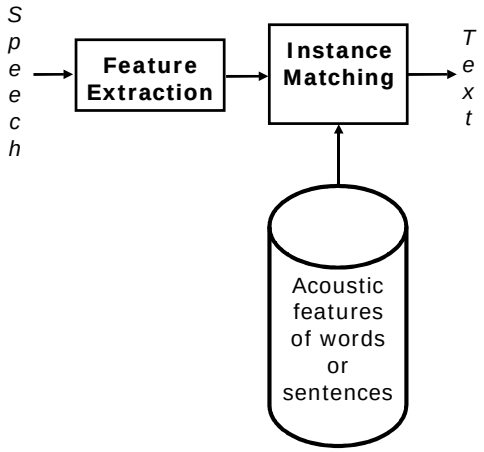


Fig. 4.1 Reconnaissance "par l'exemple" (DTW).

4.3.2 Reconnaissance par modélisation d'unités de parole

Des que l'on cherche à concevoir un système réellement multilocuteurs, à plus grand vocabulaire, et s'adaptant facilement à une application, il devient nécessaire de mener la reconnaissance sur base d'*unités de parole* de plus petite taille (typiquement les phonèmes). On ne se contente plus alors d'exemples de ces unités, mais on cherche plutôt à en déduire un *modèle* (un modèle par unité), qui sera applicable pour n'importe quelle voix.

Le formalisme de reconnaissance de la parole est alors souvent décomposé en plusieurs modules, généralement au nombre de quatre:

1. Un *module de traitement du signal* et d'analyse acoustique qui transforme le signal de parole en une séquence de *vecteurs acoustiques* (typiquement : un vecteur de coefficients LPC ou assimilés toutes les 10 ms).
2. Un *module acoustique* qui peut produire une ou plusieurs hypothèses phonétiques pour chaque segment de parole de 10 ms (c.-à-d. pour chaque vecteur acoustique), associées en général à une probabilité. Ce générateur d'hypothèse locales est généralement basé sur des *modèles statistiques* d'unités élémentaires de parole (typiquement des phonèmes) qui sont *entraînés* sur une grande quantité de données de parole (par exemple, enregistrement de nombreuses phrases) contenant plusieurs fois les différentes unités de parole dans plusieurs contextes différents. Ces modèles statistiques sont le plus souvent constitués de lois statistiques paramétriques dont on ajuste les paramètres pour « coller » au mieux aux données, ou de réseaux de neurones artificiels *(ANN : Artificial Neural Networks).* Un tel générateur d'étiquettes phonétiques intègre toujours un *module d'alignement temporel* (pattern matching, en anglais) qui transforme les hypothèses locales (prises sur chaque vecteur acoustique indépendamment) en une décision plus globale (prise en considérant les vecteurs environnants). Ceci se fait le plus souvent via des modèles de Markov cachés *(HMM pour ``Hidden Markov Model'', en anglais)*. L'ensemble (lois statistiques paramétriques ou réseau de neurones + HMM) constitue le *modèle acoustique* sous-jacent à un reconnaisseur de parole.
3. Un *module lexical* qui interagit avec le module d'alignement temporel pour forcer le reconnaisseur à ne reconnaître que des mots existants effectivement dans la langue considérée. Un tel module lexical embarque en général des *modèles des mots* de la langue (les modèles de base étant de simples dictionnaires phonétiques ; les plus complexes sont de véritables *automates probabilistes*, capables d'associer une probabilité à chaque prononciation possible d'un mot).
4. Un *module syntaxique* qui interagit avec le module d'alignement temporel pour forcer le reconnaisseur à intégrer des contraintes syntaxiques, voire sémantiques. Les connaissances syntaxiques sont généralement formalisés dans un *modèle de la langue*, qui associe une probabilité à toute suite de mots présents dans le lexique.

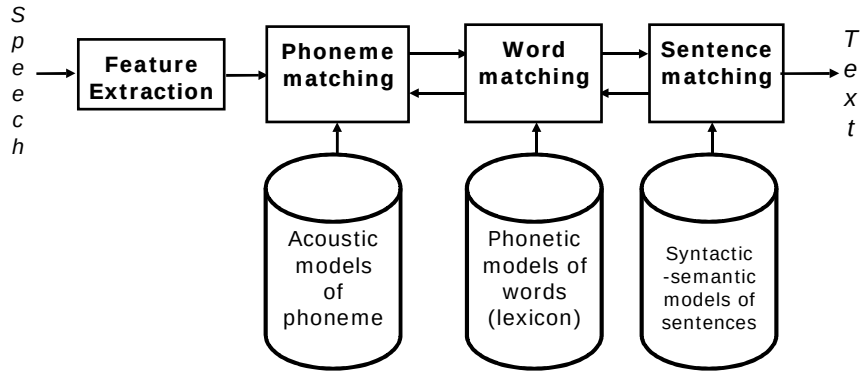


Fig. 4.2 Reconnaissance “par modélisation d’unités acoustiques”.

Les performances obtenues par de tels systèmes, même si elles sont encore loin des performances humaines, permettent aujourd’hui d’envisager sereinement l’intégration de fonctionnalités de reconnaissance vocale dans des applications pratiques.

Type	Task	Mode	Vocabulary	error rate
Isolated words	Equiprobable words	Sp. Depdt	10 digits	0%
		Sp. Indepdt	39 ascii	4.5%
		Sp. Indepdt	1109 basic English	4.3%
		Sp. Indepdt	10 digits	0.1%
		Sp. Indepdt	39 ascii	7.0%
		Sp. Indepdt	1218 names	4.7%
Connected words	Sequence of digits	Sp. Depdt	10 digits	0.1%
	id.	Sp. Indepdt	11 digits	0.2%
	Flight reservation	Sp. Depdt	129 words	0.1%
Continuous speech	Ressource management (perplexity 60)	Sp. Indepdt	991 words	3.0%
	Airline travel information system (perplexity 25)	Sp. Indepdt	1800 words	3.0%
	Wall street journal (perplexity 145)	Sp. Indepdt	20000 words	12.0%

Fig. 4.3 Etat de l'art des taux de reconnaissance (1997).

4.4 Formalisation mathématique – Approche Bayésienne de la reconnaissance de la parole

Nous abordons maintenant le problème de la formalisation mathématique, dite *Bayésienne*¸ du processus de reconnaissance de la parole par modélisation d'unités acoustiques.

Appelons $\boldsymbol{X}=\{x_1,x_2,\ldots,x_N\}$ la suite des vecteurs acoustiques du mot à reconnaître (avec $x_l=[x_{l1},x_{l2},\ldots, x_{ld}]^T$: un vecteur acoustique, par exemple de type LPC), et $\boldsymbol{M_1},\boldsymbol{M_2},\ldots, \boldsymbol{M_J}$ les *modèles statistiques* (que nous supposons exister) des J phrases que nous supposerons pouvoir reconnaître. Le problème de la reconnaissance de la parole peut alors s'exprimer sous la forme tout simple suivante : on reconnaîtra la phrase $\boldsymbol{M_{best}}$ qui maximisera $P(\boldsymbol{M_j}|\boldsymbol{X})$, où $P(\boldsymbol{M_j}|\boldsymbol{X})$ est la probabilité que $\boldsymbol{X}$ corresponde à la phase j.

L'estimation de cette probabilité nécessiterait toute fois de disposer d'un nombre tellement grand de prononciations de toutes les phrases que la prononciation $\boldsymbol{X}$ s'y trouverait elle même un grand nombre de fois. On pourrait alors estimer la probabilité par simple comptage : nombre de fois que $\boldsymbol{X}$ a été prononcé par qqn qui voulait produire la phrase j / nombre de fois total que $\boldsymbol{X}$ a été prononcé.

Il est clair que, vu la variabilité inhérente à la parole, cette situation est impossible.

On a alors recours à la formule de Bayès :

$$P(\mathbf{M}_j \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \mathbf{M}_j).P(\mathbf{M}_j)}{P(\mathbf{X})}$$

qui permet d'exprimer la probabilité inconnue sous la forme d'un rapport dont seul le numérateur est fonction de j. La maximisation sur j de ce rapport peut donc se faire sur son seul numérateur. On reconnaîtra donc la phrase $\boldsymbol{M_{best}}$ qui maximisera le produit $P(\boldsymbol{X}|\boldsymbol{M_j})\ .\ P(\boldsymbol{M_j})$. Le premier facteur intervenant dans ce produit est appelé *vraisemblance* de $\boldsymbol{X}$ selon le modèle $\boldsymbol{M_{best}}$. Il s'agit en effet de la probabilité avec laquelle le modèle $\boldsymbol{M_{best}}$ rend compte de $\boldsymbol{X}$. Le second terme est la probabilité *a priori* de la phrase modélisée par $\boldsymbol{M_{best}}$.

On peut aller plus loin encore, si on considère que la phrase modélisée par $\boldsymbol{M_{best}}$ peut elle-même être généralement prononcée de plusieurs façons phonétiques possibles (par exemple, la phrase « je suis ici » peut être prononcée [ʒəs izisi] par certains, [ʒəsuiisi] par d'autres, ou [ʃuiisi] par d'autres encore). Ces prononciations, que nous noterons $\boldsymbol{P_l}$ (l=1…L) ne sont généralement pas équiprobables. Elles sont par contre exclusives. On peut par conséquent décomposer $P(\boldsymbol{X}|\boldsymbol{M_j})$:

$$P(\boldsymbol{X}|\boldsymbol{M_j}) = P(\boldsymbol{X}|\boldsymbol{P_1})P(\boldsymbol{P_1}|\boldsymbol{M_j}) + P(\boldsymbol{X}|\boldsymbol{P_2})P(\boldsymbol{P_2}|\boldsymbol{M_j}) + \ldots + P(\boldsymbol{X}|\boldsymbol{P_L})\,P(\boldsymbol{P_L}|\boldsymbol{M_j})$$

On obtient ainsi finalement une expression de $P(\boldsymbol{M_j}|\boldsymbol{X})$ qui fait intervenir :

- $P(\boldsymbol{M_j})$, qui sera estimée à l'aide de ce que l'on appelle un « modèle de la langue ». C'est la probabilité qu'un locuteur devant prononcer une phrase quelconque dans sa langue prononce justement $\boldsymbol{M_j}$.
- $P(\boldsymbol{P_i}|\boldsymbol{M_j})$, qui sera estimée à l'aide de ce que l'on appelle un « modèle phonétique». C'est la probabilité qu'un locuteur devant prononcer la phrase $\boldsymbol{M_j}$ la prononce justement sous la forme phonétique $\boldsymbol{P_i}$.
- $P(\boldsymbol{X}|\boldsymbol{P_j})$, qui sera estimée à l'aide de ce que l'on appelle un « modèle acoustique». C'est la probabilité qu'un locuteur devant prononcer la suite phonétique $\boldsymbol{P_i}$ la prononce justement sous la forme acoustique $\boldsymbol{X}$.

Chacune de ces probabilités peut être estimée à partir d'une grande bases de données de parole, moyennant certaines hypothèses (et donc certaines erreurs dans l'estimation des probabilités).

4.5 Pistes à suivre

Même si on obtient aujourd'hui des bons scores de reconnaissance (voir Fig. 4.3), le problème est loin d'être résolu. Les chiffres mentionnée à la figure 4.3 sont en effet ceux obtenus dans des conditions de laboratoire. Il est frappant de constater que, dès lors que l'on place les reconnaisseurs dans des conditions réelles de bruit ambiant et de grande variabilité des conditions d'enregistrement (position par rapport au micro, par exemple), les taux de reconnaissance s'effondrent littéralement (-30%). Ce problème de *robustesse* est certainement un des grands défis de la reconnaissance vocale pour les années à venir.

Un problème connexe est celui de la séparation de signaux vocaux superposés. Nous sommes tous capables de focaliser notre attention sur une conversation dans un cocktail, même si le signal qui parvient à nos oreilles est constitué de la superposition d'un grand nombre de conversations indépendantes. Aucun algorithme n'a pu être développé a ce jour permettant d'automatiser cette capacité étonnante de notre cerveau. Il s'agit pourtant souvent d'une faculté intervenant de façon importante dans notre capacité de reconnaissance de parole.

On constate par ailleurs généralement que les modèles (syntaxiques) de la langue sont responsables des derniers 10% dans les scores obtenus, ce qui montre bien leur importance. On pouvait évidemment s'y attendre, vu l'importance que peut prendre, pour les êtres humains que nous sommes, la connaissance de la syntaxe d'une langue en vue de sa bonne compréhension. Or précisément, ces mêmes modèles de la langue sont encore très rudimentaires. Et ce n'est pas parce que les chercheurs sont à court d'idées dans ce domaine, mais bien plutôt parce que tous les essais effectués pour rendre le modèle syntaxique plus complet conduisent à une moins bonne estimation de ses paramètres, et donc finalement à une moins bonne estimation de $P(\boldsymbol{M_j})$.

CONCLUSION

Depuis une décennie, les techniques de traitement de la parole ont connu plusieurs grandes révolutions.

La première, et celle qui touche pour l'instant de loin le plus d'utilisateurs, est celle de la téléphonie mobile : une proportion grandissante de la population transporte avec elle un ordinateur de poche spécialisé dans l'analyse-synthèse LPC. Les algorithmes de codage sont par ailleurs également utilisés dans les boîtes vocales : nos paroles y sont stockées sous la forme de suites de vecteurs de paramètres LPC. Le marché du codage de la parole est donc à présent largement ouvert, ce qui n'est pas le cas en reconnaissance ou en synthèse.

La seconde révolution est celle des grandes bases de données de parole et de textes. Depuis 1995, sous l'égide de LDC (Language Data Consortium) aux Etats-Unis et de l'ELRA (European Language Resource Agency) en Europe, de nombreux laboratoires de recherche (publics et privés) mettent en commun leurs ressources. Il en résulte un foisonnement de données propices à l'établissement de modèles, tant numériques que symboliques, de la parole. Les développements récents reconnaissance, et plus encore en synthèse, en sont en grande partie la conséquence logique.

Une troisième révolution, liée à la précédente, est celle des outils d'ingénierie pure (HMMs, ANNs, Synthèse par sélection d'unités dans une grande base de données), qui tend à supplanter de plus en plus l'expertise humaine (reconnaissance analytique, synthèse par règles), laquelle intervient plutôt au second plan, en permettant d'affiner les résultats.

Enfin, une dernière révolution se prépare : celle qui verra naître des machines dont plus personne ne pourra affirmer avec certitude qu'elles en sont. Aujourd'hui déjà, la qualité des algorithmes de synthèse vocale permet aux synthétiseurs de passer avec succès le fameux « test de Turing », inventé par le mathématicien anglais Alan Turing dans les années 40 pour mesurer le degré d' « intelligence » d'une machine : en vérifiant combien de temps un expérimentateur interagissant « en aveugle » avec cette machine peut rester persuadé d'avoir affaire à un être humain. Les reconnaisseurs sont eux-mêmes

prêts tout prêts à tromper notre intelligence, et ils ne manqueront pas de la faire, dès lors que l'on aura amélioré leurs capacités de robustesse.

Il n'en reste pas moins que, alors que nos ordinateurs pourront nous parler et reconnaître ce que nous leur dirons, ils n'en seront pas pour autant capables de *comprendre* nos parole. C'est là un tout autre domaine, dont nous ne connaissons encore que les tout premiers balbutiements.