

« Je parle, donc je suis ? »

## Un bilan des développements récents en traitement automatique de la parole

T. Dutoit

Faculté Polytechnique de Mons

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.

La parole est en effet un des seuls à être à la fois *produit* et *perçu* instantanément par le cerveau. 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 influencer sur sa production par le biais de ce bouclage, mais aussi et surtout

sa *compréhension*<sup>1</sup>. On ne parle que dans la mesure où l'on s'entend et où l'on se comprend soi-même; la complexité du signal qui en résulte s'en ressent forcément<sup>2</sup>.

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 **reconnaisseurs** ont pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse.
- Les **synthétiseurs** ont quant à eux la fonction inverse de celle des reconnaisseurs de parole : ils produisent de la parole artificielle.
- 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"*.

---

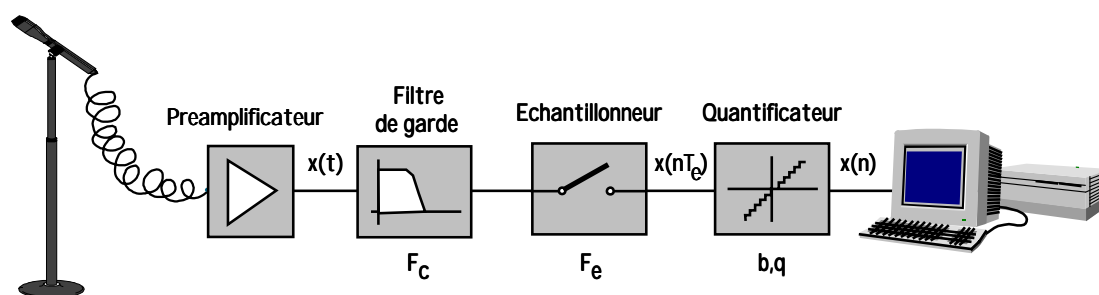
<sup>1</sup> Ceci accroît encore la différence entre la parole et, par exemple, l'image : alors que la compréhension de l'image est en principe également accessible à tous, la compréhension de la parole est le résultat d'un apprentissage socio-culturel lié à une *communauté linguistique*.

<sup>2</sup> 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.

## Qu'est-ce que la parole ?

L'information portée par le signal de parole peut être analysée de bien des façons. Nous en rappellerons simplement ici les aspects acoustiques.

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La **phonétique acoustique** 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**.



**Fig. 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  $q$ .

L'échantillonnage (Fig. 1) transforme le signal à temps continu  $x(t)$  en signal à temps discret  $x(n)$  défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage; celle-ci est elle-même l'inverse de la fréquence d'échantillonnage. Pour ce qui concerne le signal vocal, le choix de cette fréquence d'échantillonnage résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz. Il faut donc en principe choisir une fréquence égale à 24 kHz au moins pour satisfaire raisonnablement au théorème de Shannon. 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. Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit une fréquence d'échantillonnage égale à 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.

Parmi le continuum des valeurs possibles pour les échantillons  $x(n)$ , 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  $x(n)$ . Une quantification de bonne qualité requiert en général 16 bits.

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 sur une ligne RNIS exige un débit de  $8 \text{ kHz} \times 8 \text{ bits} = 64 \text{ kb/s}$ ; la transmission ou l'enregistrement d'un signal audio exige en principe un débit de l'ordre de  $48 \text{ kHz} \times 16 \text{ bits} = 768 \text{ kb/s}$  (à multiplier par deux pour un signal stéréophonique)<sup>3</sup>.

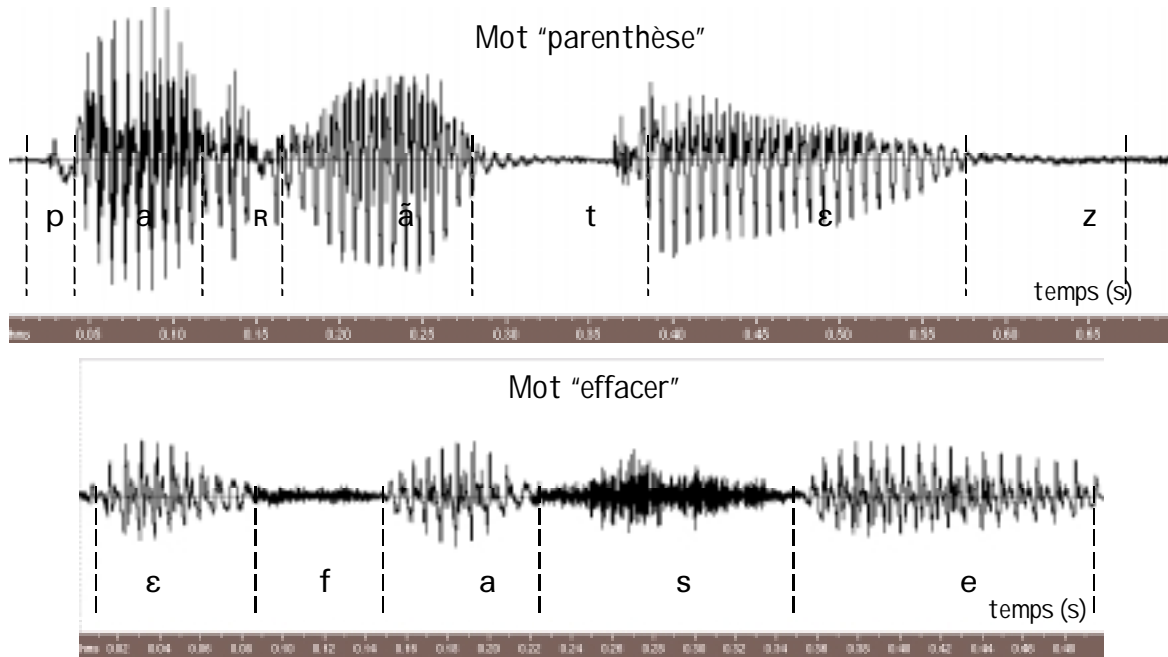
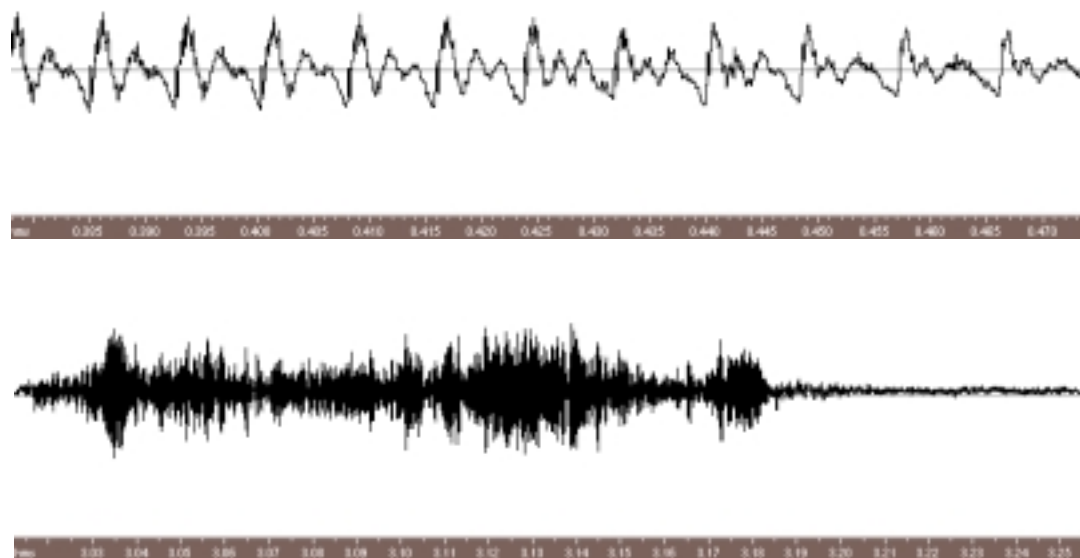


Fig. 2 Audiogramme de signaux de parole.

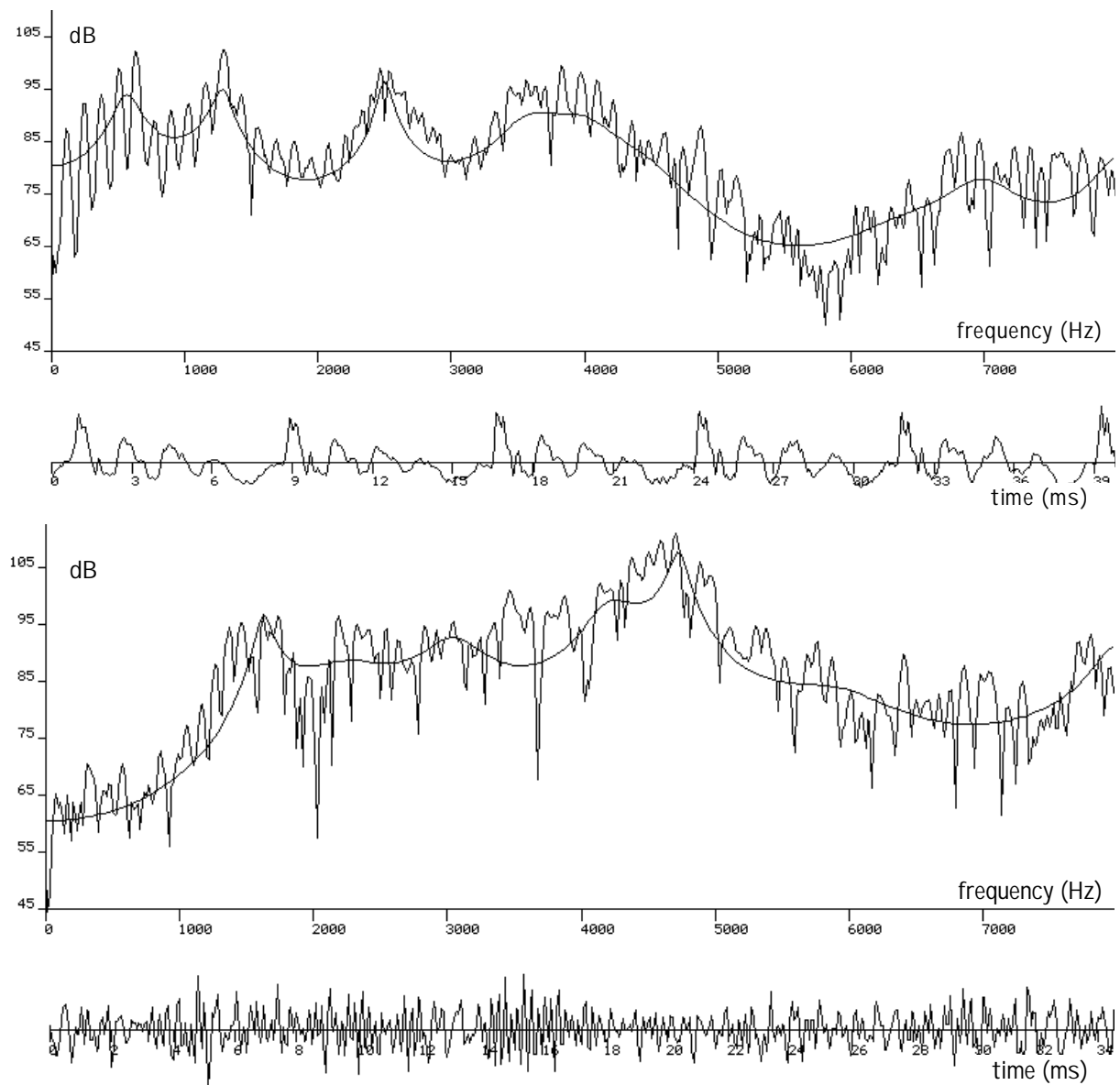
<sup>3</sup> La redondance naturelle du signal vocal permet de réduire le débit binaire dans une très large mesure, au prix d'un traitement plus ou moins complexe et au risque d'une certaine dégradation de la qualité de la représentation. Cette question sera abordée plus loin.



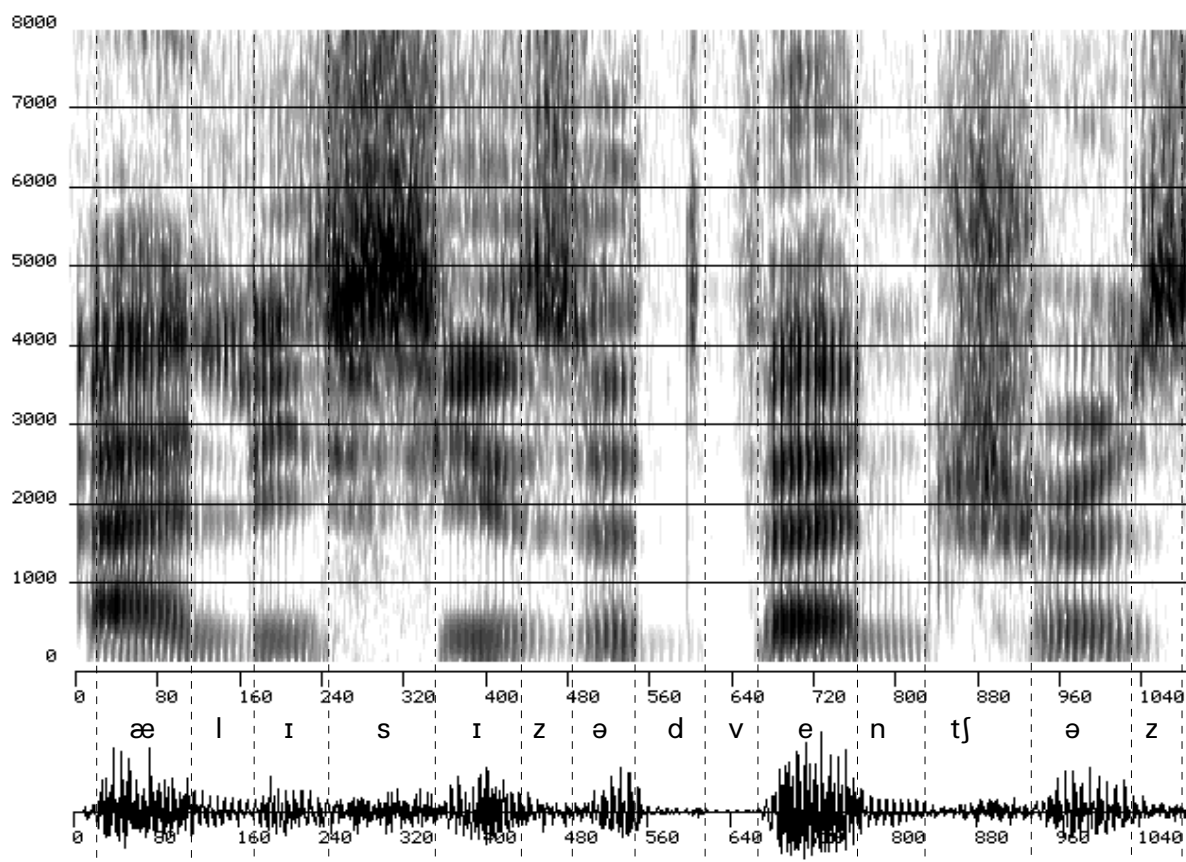
**Fig. 3** Exemples de son voisé (haut) et non-voisé (bas).

La figure 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 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.

La transformée de Fourier à court terme est obtenue en extrayant de l'audiogramme une trentaine de millisecondes de signal vocal et en effectuant une transformée de Fourier sur ces échantillons. Le résultat de cette transformation mathématique est souvent présenté dans un graphique qui donne, en fonction de la fréquence, l'amplitude des composantes présentes dans le signal analysé.



**Fig. 4** Evolution temporelle (en haut) et transformée de Fourier discrète (en bas) du [a] et du [ʃ] de 'baluchon' (tranche de 30 ms).



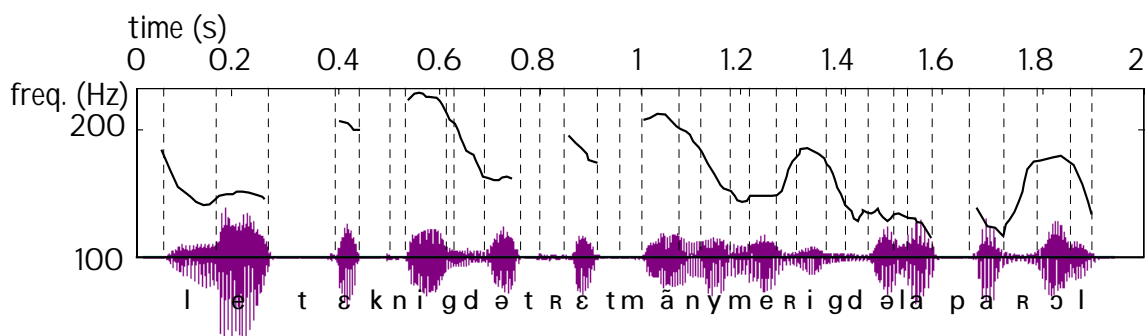
**Fig. 5** Spectrogramme et audiogramme de la phrase anglaise 'Alice's adventures', échantillonnée à 11.25 kHz.

La figure 4 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. 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. Le spectre d'un signal de type voisé possède en général plus de composantes en basse fréquence qu'en haute fréquence. Par contre, le spectre d'un signal non voisé présente une amplitude plus importante en haute fréquence (ce qui correspond à la perception que nous en avons : les fricatives non-voisées [f,s,j] sont des sons plus « aigus » que les voyelles).

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. Ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des

formants. On reconnaît sans peine, à la figure 5, les formants caractéristiques des voyelles, et les composantes à plus hautes fréquences caractéristiques des consonnes. La position et l'évolution des formants est caractéristique des sons produits. La seule lecture d'un spectrogramme (sans l'écoute du signal correspondant) permet d'ailleurs à l'œil expérimenté de certains phonéticiens de retrouver le contenu du message parlé. Cette propriété n'est évidemment pas vraie pour l'audiogramme, qui renseigne peu sur le timbre des séquences sonores produites. C'est donc bien que le spectrogramme présente sous une forme simple l'essentiel de l'information portée par le signal vocal.

Notons pour terminer qu'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 ou *pitch*. La figure 6 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.



**Fig. 6** 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..

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. 7).

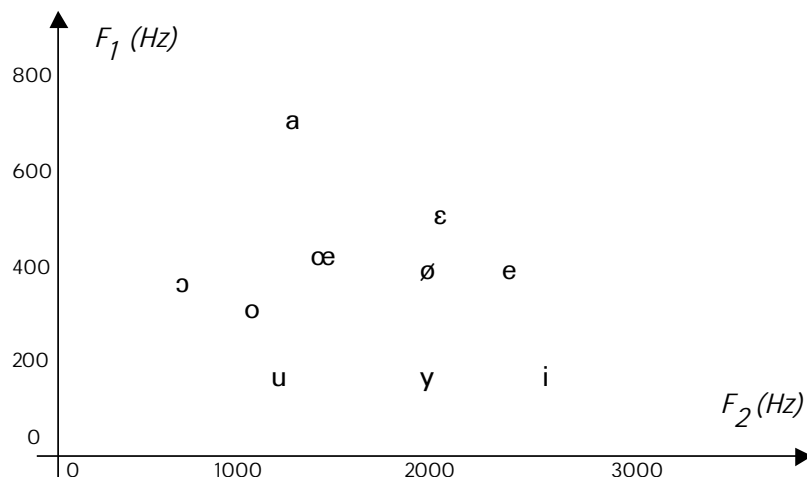
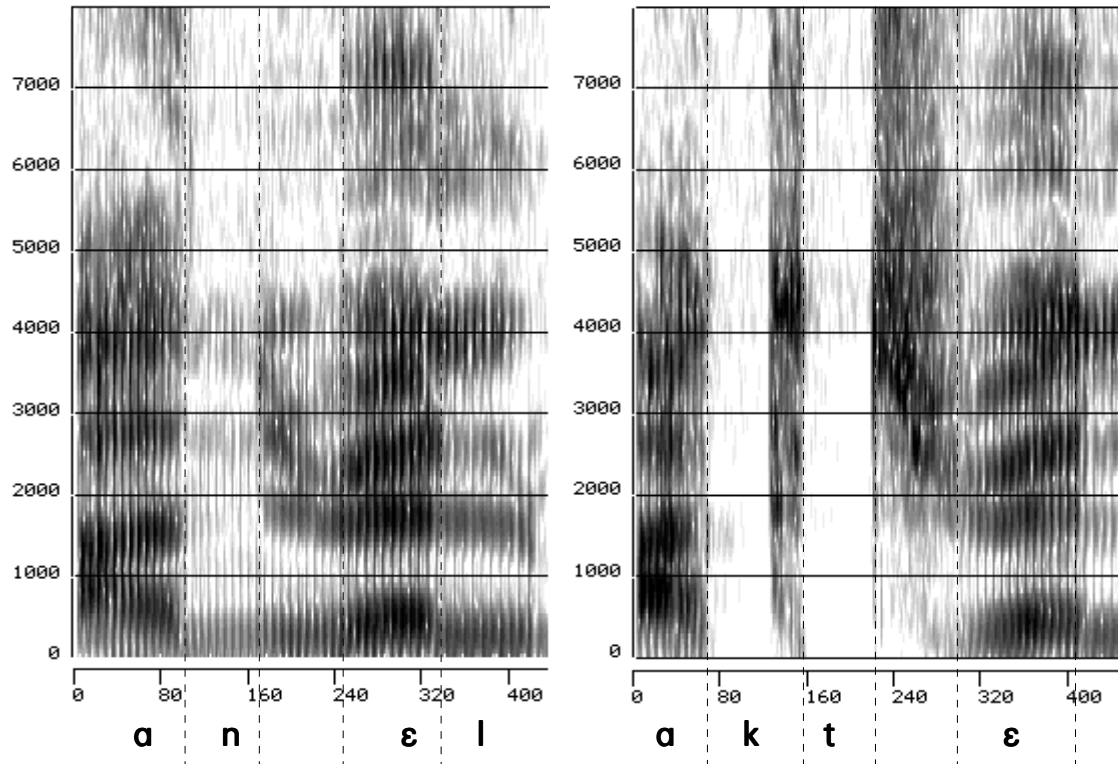


Fig. 7 Représentation des voyelles dans le plan F1-F2

On observe en pratique un certain recouvrement dans les zones formantiques correspondant à chaque voyelle. D'une manière plus générale, d'ailleurs, voyelles et consonnes apparaissent 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 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 **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 (on en trouve un exemple frappant à la figure 8). **Ces phénomènes de coarticulation sont 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.**



**Fig. 8** Un cas d'assimilation de sonorité (coarticulation affectant le voisement d'une sonore). A gauche, le début du mot '*annuellement*', dans lequel [ ] est placé dans un contexte voisé. A droite, le début de '*actuellement*' : [ ] est totalement dévoisé à cause de la plosive sourde qui précède.

## Le modèle prédictif Linéaire du signal vocal

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. 9).

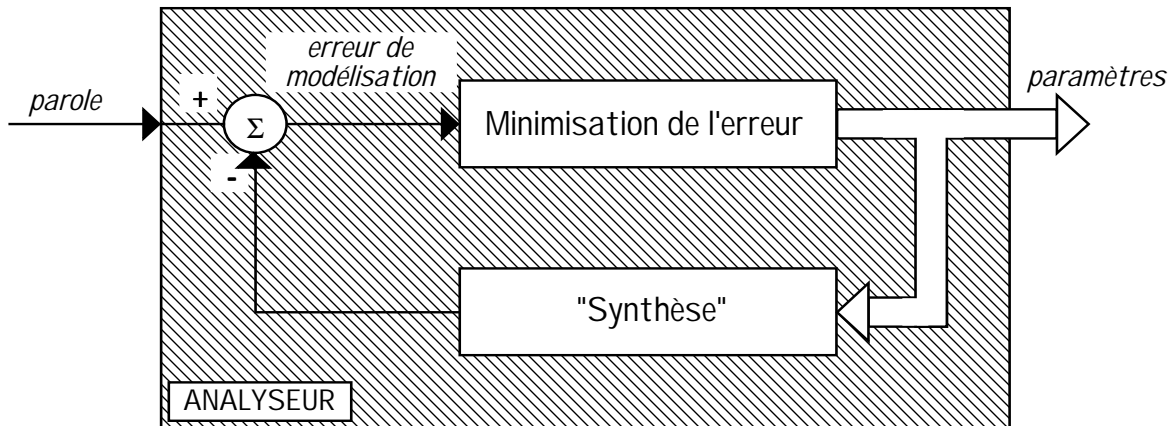


Fig. 9 Schéma de principe d'un analyseur de parole. En pratique, l'étape de synthèse peut être implicite.

S'il existe de nombreux modèles de parole, il en est un que l'on retrouve partout, et dans un nombre croissant d'appareils « grand-public » : le modèle prédictif linéaire (LPC : Linear Predictive Coding).

De la même façon qu'un signal de parole réel est produit par le passage, à travers le filtre que constitue notre conduit vocal, d'un signal d'excitation créé par les poumons et les cordes vocales, ce même signal de parole peut être modélisé par le passage d'un signal d'excitation numérique à travers un filtre numérique récuratif. Le signal d'excitation sera tantôt une suite d'impulsions numériques (qui serviront à simuler les impulsions de débit créées par les cordes vocales) tantôt du bruit numérique (qui reproduira le souffle poussé par les poumons).

Ce modèle est appelé « prédictif linéaire » en raison du fait qu'il correspond à une régression linéaire très simple entre le signal d'excitation et le signal vocale produit. Les coefficients de cette régression linéaire sont les coefficients du filtre numérique récuratif.

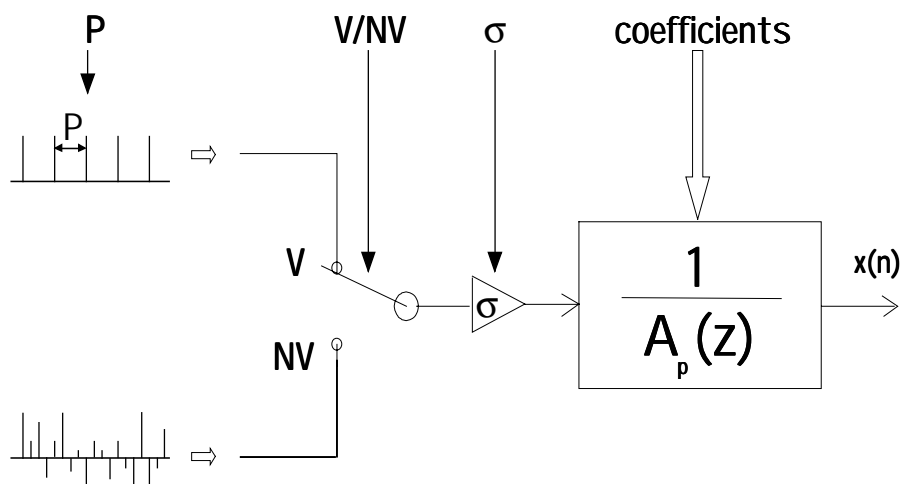


Fig. 10 Le modèle auto-régressif.

Les paramètres<sup>4</sup> du modèle LPC sont tout simplement : la période du train d'impulsions (sons voisés uniquement), la position de l'interrupteur Voisé/NonVoisé (V/NV), le gain de l'amplificateur  $\sigma$ , et les coefficients du filtre numérique de synthèse.

Le problème de l'estimation d'un modèle LPC, souvent appelée *analyse LPC* revient à déterminer les coefficients du filtre connaissant 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* :

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

avec

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

$$\mathbf{a} = [a_1, a_2, \dots, a_p]^T$$

$$\Phi = \begin{pmatrix} \phi_{xx}(0) & \phi_{xx}(1) & \phi_{xx}(2) & \dots & \phi_{xx}(p-1) \\ \phi_{xx}(1) & \phi_{xx}(0) & \phi_{xx}(1) & \dots & \phi_{xx}(p-2) \\ \phi_{xx}(2) & \phi_{xx}(1) & \phi_{xx}(0) & \dots & \phi_{xx}(p-3) \\ \vdots & \vdots & \vdots & & \vdots \\ \phi_{xx}(p-1) & \phi_{xx}(p-2) & \phi_{xx}(p-3) & \dots & \phi_{xx}(0) \end{pmatrix}$$

Les coefficients de la matrice et du terme indépendant du système sont les coefficients d'autocorrélation du signal à modéliser. Les inconnues sont les coefficients du filtre de synthèse, dont le nombre,  $p$ , est typiquement de l'ordre de 10. Le signal de parole étant fortement non-stationnaire, ce type de modélisation ne reste guère valable plus d'une dizaine de millisecondes. On retiendra donc que **l'analyse LPC d'un signal de parole implique la résolution d'un système de 10 d'équations à 10 inconnues toutes les 10 ms.**

## 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**. Le 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

---

<sup>4</sup> On trouvera sur le site web du cours de traitement de parole dispensé à la Faculté Polytechnique de Mons (<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.

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 millisecondes 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 son visage, lequel serait reproduit en image de synthèse côté récepteur.

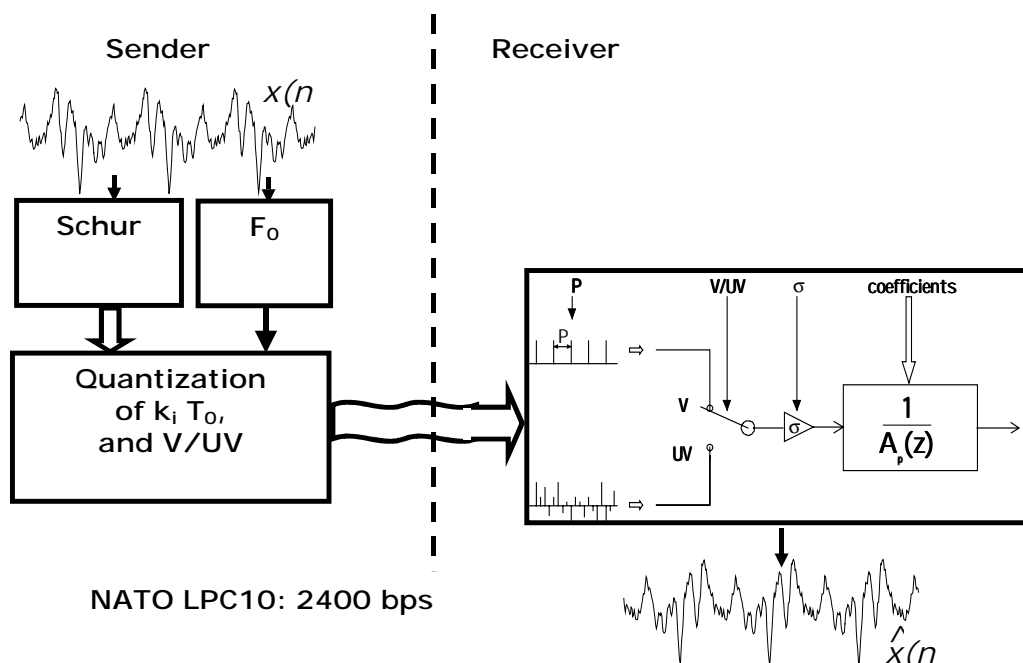


Fig. 11 Transmission de parole basée sur le codage LPC

Les débits de transmission obtenus avec ce type de modèle (et ses perfectionnements) sont respectivement de 2400 bits par seconde pour la transmission de voix par satellite, et de 13400 bits par seconde pour la transmission par GSM<sup>5</sup>. Ces chiffres sont à comparer aux 64000 bits par secondes pour la téléphonie numérique RNIS et aux 33600 bits par seconde que nos modems peuvent envoyer sur nos lignes téléphoniques analogiques. Le GSM

<sup>5</sup> Ceci explique en passant pourquoi la transmission de données numériques est limitée à 9600 bits par secondes avec un GSM.

n'aurait tout simplement pas pu voir le jour sans les efforts consentis pour parvenir à ces débits.

## Synthèse de la parole

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'égaliser 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.

La suite des opérations à effectuer pour obtenir la lecture automatique d'une phrase ne se conforme par conséquent que de loin au schéma fonctionnel tout naturellement adopté par le cerveau. La parole naturelle est en effet 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. Or, 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 de synthèse « du texte à la parole » (TTS : text-to-speech) développés 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 !*"

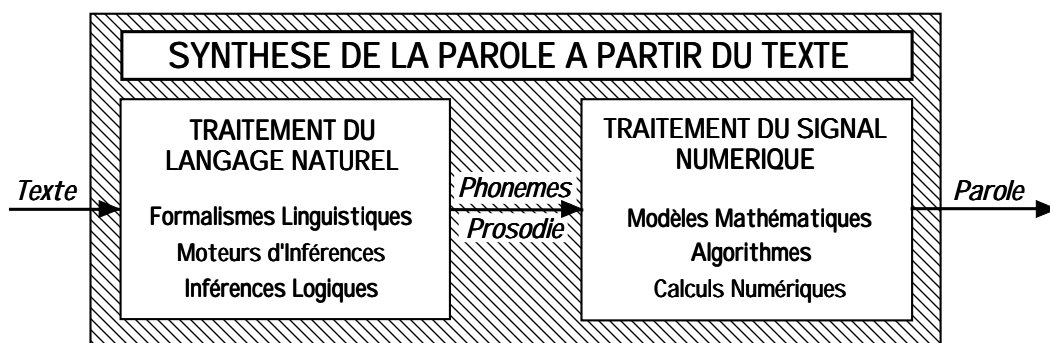


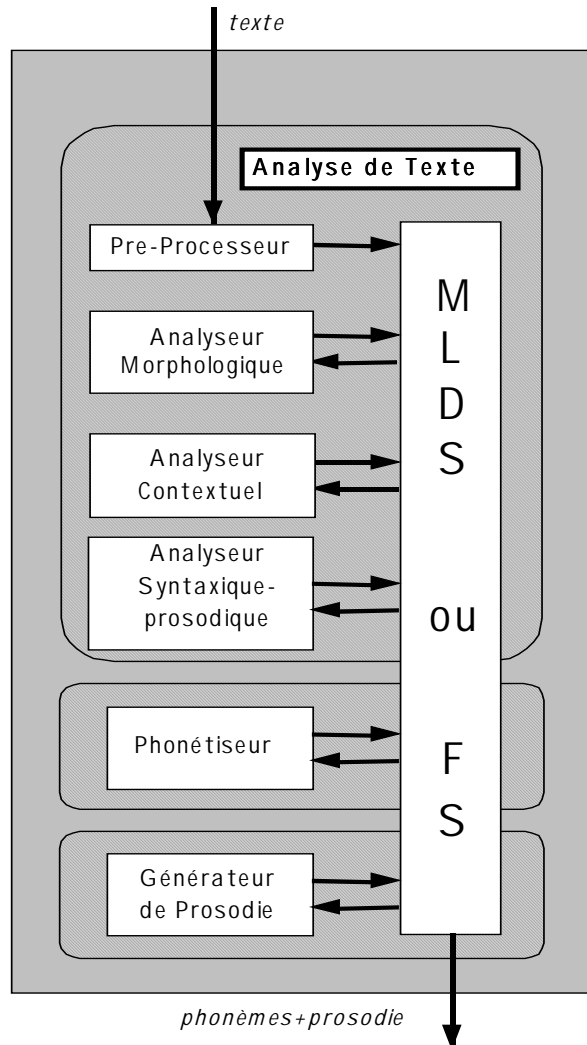
Fig. 12 Diagramme fonctionnel d'un système de synthèse TTS.

La figure 12 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.

L'organisation générale des opérations de traitement du langage réalisées par le synthétiseur est donnée à la figure 13. On y remarque immédiatement, outre la présence attendue des modules de *phonétisation automatique* et de *génération de la prosodie*<sup>6</sup>, l'importance du module d'*analyse morpho-syntaxique*, indispensable à la bonne prononciation du texte à lire. Il est en effet nécessaire de connaître la nature des mots pour les prononcer correctement (ex : « les poules du couvent couvent », « le président et ses associés président », « lest est à l'est », etc.). Il est par ailleurs nécessaire de déterminer les groupements de mots qui constituent à l'oral ce que l'on appelle un « groupe intonatif », sur lequel vient se placer un « mouvement intonatif » complet.

---

<sup>6</sup> Le terme « prosodie » désigne collectivement l'intonation, le rythme, et l'intensité du signal de parole, dont les valeurs évoluent au cours du temps d'une manière propre à la langue et au locuteur.



**Fig. 13** Le module de traitement du langage naturel d'un système de conversion texte-parole.

Le module d'analyse morfo-syntaxique de la figure 13 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 données 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 clair que l'ensemble des ces problèmes, qui relèvent de la *linguistique informatique*, sont toujours à l'étude. Un des plus cruciaux est sans aucun doute celui qui est lié à la modélisation de l'intonation humaine. Nous possédons tous, en effet, une connaissance implicite de l'intonation que nous appliquons sur nos parole. Bien peu de spécialiste sont cependant capables de fournir un modèle réaliste de ce processus.

Lorsqu'un texte a été phonétisé et que l'on a déterminé avec quelle intonation et quel rythme on désirait le faire prononcer, il reste évidemment à produire le signal de parole correspondant. Cette étape est du ressort du *synthétiseur* proprement dit.

Les premiers synthétiseurs électriques faisaient pour ce faire appel à l'expertise de phonéticiens, dont le rôle était de prédéterminer, pour une suite de sons donnés, l'allure du spectrogramme correspondant. Ce spectrogramme était alors transformé en parole à l'aide d'un ensemble de générateurs et de filtres électriques. En pratique, le spectrogramme était construit à l'aide de règles (établies par des experts) déterminant l'évolution des formants au cours de temps. C'est ce qui a valu à cette technique le nom de *synthèse par règles*. Elle a connu un franc succès jusque dans les années 80, et est parfois encore exploitée commercialement de nos jours.

Les années 80 ont vu le développement d'une technique de synthèse radicalement différente, basée sur la mise bout à bout d'unités vocales découpées dans un enregistrement de parole humaine. Ces techniques, dites de *synthèse par concaténation*, utilisent le plus souvent des *diphones* comme unités de base. Un diphone est une unité de parole qui commence au milieu d'un phonème et se termine au milieu du phonème suivant. La figure 14 donne un exemple des étapes nécessaires à la production du mot « dog » selon ce principe. On commence par déterminer quels diphones seront nécessaires. Ces diphones sont ensuite extraits d'une base de données (préalablement constituée, et contenant un exemple de chacun des diphones de la langue à synthétiser). Les diphones ainsi obtenus ne correspondent jamais exactement à la commande du synthétiseur : ils possèdent leur intonation et leur durée propre, qui n'est pas en général celle que l'on cherche à produire. Il faut par conséquent modifier ces caractéristiques avant de pouvoir procéder à la concaténation proprement dite. C'est le rôle du module de *modification de la prosodie*. Enfin, la simple mise bout à bout d'unités de parole extraites des contextes différent ne produit en général pas de la parole continue de bonne qualité. Il est alors nécessaire de procéder à un « lissage » des extrémités des diphones, afin d'assurer une bonne continuité des formants aux jointures. Cette opération porte le nom de *lissage spectral*.

La modification de la prosodie et le lissage spectral de signaux de parole sont des tâches difficiles à réaliser sans qu'il en résulte des dégradation dans la qualité du signal. Une bonne partie des recherches menées en synthèse vocale durant les

années 80 et 90 a précisément porté sur la mise au point de méthodes d'analyse-synthèse de parole permettant de résoudre ces problèmes avec plus ou moins de succès. Parmi les modèles les plus utilisés, on trouve la synthèse LPC (AT&T Bell Labs, USA) , la synthèse PSOLA (France Télécom CNET Lannion), et la **synthèse MBROLA** développée à la Faculté Polytechnique de Mons depuis 1993<sup>7</sup>.

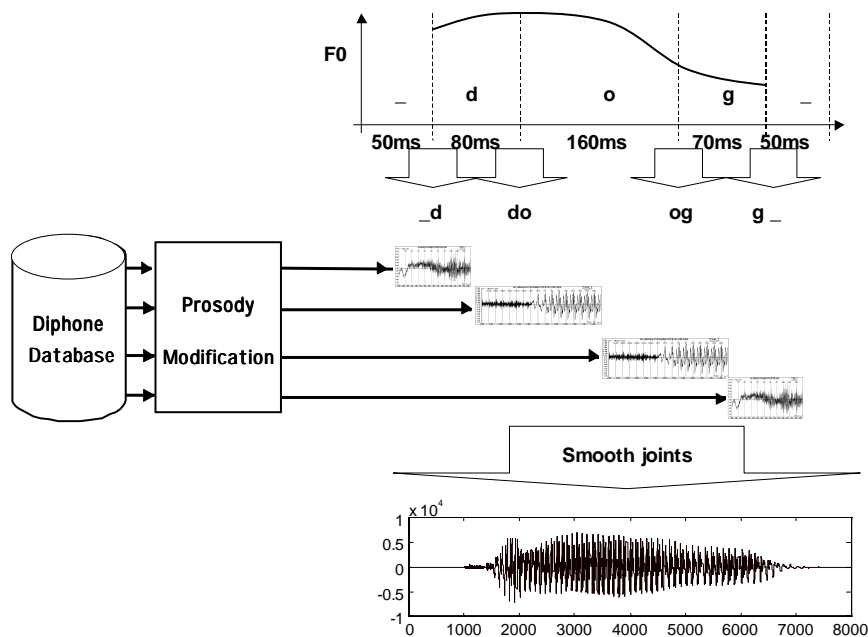


Fig. 14 Synthèse du mot « dog » par concaténation de diphones.

Plus récemment encore, la qualité des voix de synthèse a progressé de façon impressionnante, grâce aux techniques dites de **synthèse par sélection d'unités dans une grande base de données**. Le principe en est très simple : il s'agit ni plus ni moins de techniques de synthèse par concaténation dont les bases de données contiennent, non plus un seul exemple de chacun des diphones d'une langue, mais autant que possible (voir Fig. 15). La seule limite est alors le temps nécessaire à la segmentation de parole enregistrée. Ce sont ainsi plusieurs heures de paroles qui sont utilisées par le synthétiseur, qui cherche parmi toutes les unités disponibles celles qui produiront *in fine* la parole de meilleure qualité. Tout l'art du concepteur est alors d'établir les critères formels (sur base de mesure des caractéristiques acoustiques et linguistiques des diphones

<sup>7</sup> Cette technique de synthèse, brevetée internationalement depuis 1996, a directement conduit au projet MBROLA, par lequel la FPMs s'est associée à une centaine de partenaires à travers le monde entier pour la mise au point du plus grand nombre de voix de synthèse jamais réalisées. Au moment de la rédaction de cet article, le projet MBROLA propose 22 langues et 36 voix sur son site web (<http://tcts.fpms.ac.be/synthesis/mbrola/>). MBROLA a également conduit à la création de plusieurs PME pour la commercialisation des résultats ; la principale, Babel Technologies S.A. (<http://www.babeltech.com>) , est une spin-off de la FPMs établie à Mons.

disponibles) qui permettront de choisir effectivement la meilleure suite de diphone possible.

Lorsque la taille de la base de données est suffisamment grande et que le choix des unités s'effectue au mieux, les voix de synthèse produites grâce à ces techniques sont d'une qualité telle qu'il est souvent impossible de les distinguer d'une voix humaine. Il va sans dire que ces techniques suscitent actuellement un intérêt commercial très important.

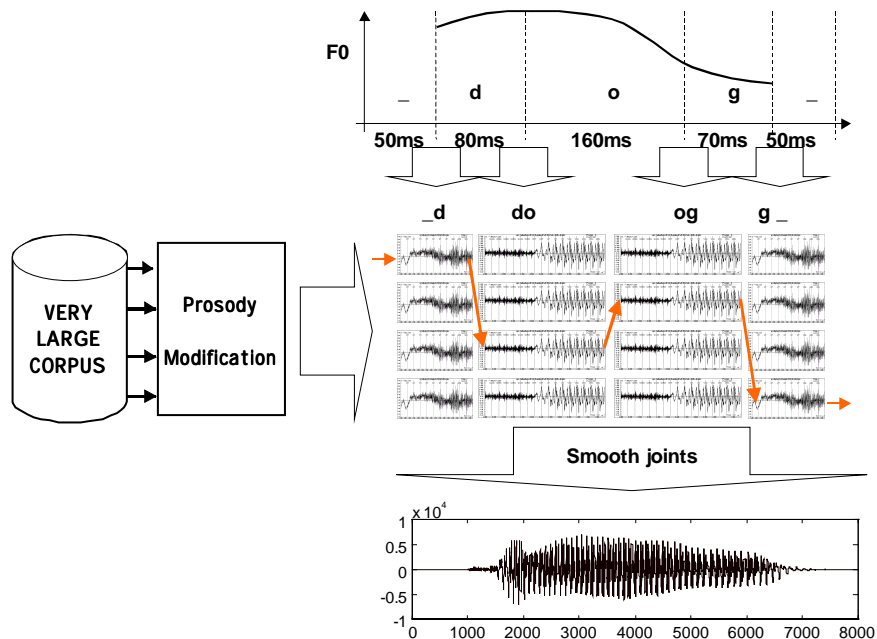


Fig. 15 Synthèse du mot « dog » par sélection d'unités dans une grande base de données.

## Reconnaissance de la parole

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).

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 qu'il nous soit donné d'étudier. 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 milieu ambiant).

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 reconnaissance 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%).

Des que l'on cherche à concevoir un système réellement multilocuteur, 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 reconnaiseur de parole.
3. Un *module lexical* qui interagit avec le module d'alignement temporel pour forcer le reconnaiseur 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 reconnaiseur 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.

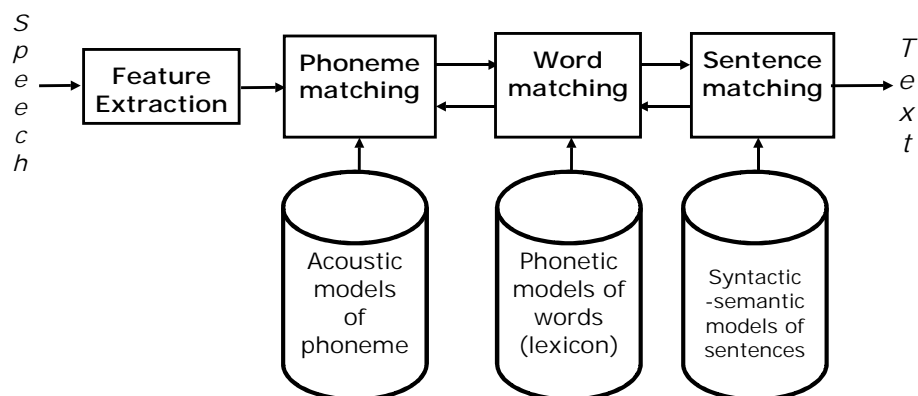


Fig. 4.2 Reconnaissance "par modélisation d'unités acoustiques".

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 (voir Table 1. ?).

Il est cependant à noter que la plupart des bons résultats avancés correspondent à des conditions de laboratoire. Il est alors 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.

Type	Tâche	Mode	Vocabulary	error rate
Mots isolés	Mots équiprobables	Dépendant du locuteur	10 chiffres	0%
	Id.	Indépendant du locuteur	Id.	0.1%
	Id.	Dépendant du locuteur	39 caractères ascii	4.5%
	Id.	Indépendant du locuteur	Id.	7.0%
	Id.	Indépendant du locuteur	1109 mots de base	4.3%
	Id.	Indépendant du locuteur	1218 noms	4.7%
Parole continue	Ressource management	Indépendant du locuteur	991 mots	3.0%
	Demande d'informations dans un environnement	Indépendant du locuteur	1800 mots	3.0%

---

un aéroport					
Lecture du Street Journal	Wall	Indépendant du locuteur	20000 mots	12.0%	

---

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

## Conclusion

En 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 souvent sans le savoir 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 encore 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 (modèles de Markov cachés, réseaux de neurones artificiels, synthèse par sélection d'unités dans une grande base de données), qui tendent 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 paroles. C'est là un tout autre domaine, dont nous ne connaissons encore que les tout premiers balbutiements.

## **Pour aller plus loin :**

Une partie du texte et des images qui constituent cet article provient de l'ouvrage suivant :

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

Nous conseillons vivement au lecteur intéressé aux aspects techniques du traitement de la parole de s'y reporter.

Il existe par ailleurs un site web qui centralise l'ensemble des informations disponibles sur la recherche, le développement, et le commercialisation de technologies et de produits vocaux : la « speech FAQ » (<http://www.speech.cs.cmu.edu/comp.speech/>). C'est incontestablement le meilleur point d'entrée dans le monde des technologies vocales sur internet.