Chapitre 1

Principes fondamentaux de la synthèse de parole

Pourquoi la synthèse de parole ?

Avant de décrire l'adaptation du synthétiseur aux différents micro-ordinateurs, nous commencerons par une explication des principes généraux de la synthèse de la parole, suivie de la description de leur mise en application dans le synthétiseur à formants MEA 8000.

En premier lieu, on peut se demander pourquoi utiliser un synthétiseur de parole, alors que d'autres méthodes de reproduction a priori plus simples et plus directes existent. Deux raisons principales y conduisent :

• Si l'on veut disposer d'un accès aléatoire et immédiat à tous les éléments composant un vocabulaire donné, les moyens électromécaniques traditionnels (bandes ou disques magnétiques) sont inapropriés. Il n'est pour s'en convaincre que de regarder la complexité et le coût d'une

réalisation telle que l'horloge parlante, dont le vocabulaire est pourtant relativement limité. Il faut donc pouvoir stocker la parole dans un dispositif entièrement statique tel qu'une mémoire à semi-conducteurs, par exemple.

Ceci nécessite donc la "numérisation" du signal vocal ; pour obtenir une qualité comparable à celle du téléphone, il faut disposer d'une bande passante de l'ordre de 4 kHz, et d'une dynamique d'au moins 40 dB. C'est ce que l'on peut atteindre par une conversion analogique/numérique sur 10 à12 bits avec échantillonnage à une fréquence de 8 kHz (théorème de Shannon). Le débit binaire résultant peut être réduit à 64 kb/s, valeur normalisée pour les transmissions téléphoniques de type M.I.C. (Modulation par Impulsions Codées).

• Le codage numérique pur (à 64 kb/s par exemple) est trop "gourmand" en mémoire : en effet, à ce débit, la mémoire vive d'un micro-ordinateur de 64 k-octets ne pourrait contenir que 8 secondes de parole, même sans aucun programme d'application !!

On a donc cherché depuis longtemps à réduire ce débit en exploitant la redondance très importante du signal vocal qui transporte un débit d'information utile de l'ordre de 70 bits par seconde, soit près de 1000 fois moins que le résultat de la numérisation du signal ; cette redondance a pour effet de rendre ce signal très résistant au bruit ambiant et aux distorsions qui peuvent affecter sa transmission.

Pour réduire le débit d'informations à transmettre, tous les synthétiseurs tirent parti du fait que la voix humaine peut seulement émettre certains sons compatibles avec le système physiologique de génération de la voix, dont nous allons étudier brièvement le fonctionnement.

Pratiquement, tous les synthétiseurs auront donc un schéma-bloc général dérivé de ce modèle physiologique ; nous étudierons plus particulièrement le modèle "à formants", dont le MEA 8000 fait partie.

La synthèse de parole à formants

La figure 1 représente une coupe de l'appareil vocal humain et sa représentation schématique simplifiée.

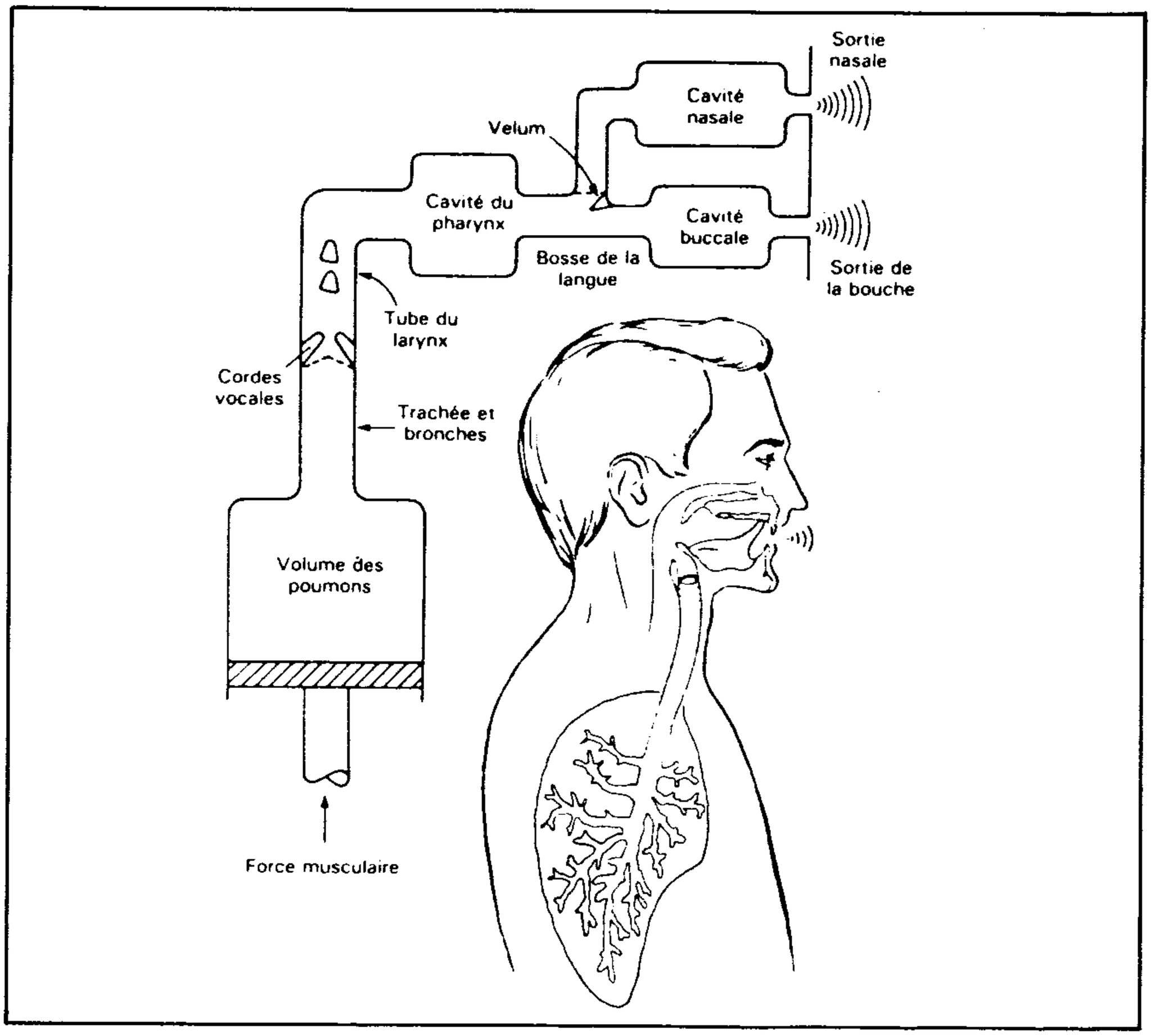

Figure 1 - Représentation simplifiée de l'appareil vocal

L'énergie qui servira à produire la voix est fournie sous forme de pression d'air par les poumons que l'on peut assimiler à une pompe. L'augmentation de la pression de l'air provoque l'ouverture des cordes vocales initialement closes. Il s'ensuit une brusque chute de pression, provoquant la fermeture des cordes vocales ; ceci entraîne une nouvelle augmentation de pression qui ouvrira de nouveau les cordes vocales, et ainsi de suite.

Ce mécanisme crée ainsi un train périodique d'impulsions de pression en dents de scie qui excite le conduit vocal situé au-dessus des cordes vocales. Les sons créés par ce processus sont dit "voisés" et correspondent à toutes les voyelles et à certaines consonnes dites sonores (b, d, g, l, m, n, r, v, z).

Le signal périodique ainsi généré est riche en harmoniques dont la décroissance est de l'ordre de 12 dB par octave ; sa fréquence fondamentale est appelée "pitch" dans la littérature anglo-saxonne et nous utiliserons par la suite ce terme ou celui de "fondamental" pour la désigner.

Il existe un autre mode de génération de sons vocaux dans lequel les cordes vocales sont toujours entr'ouvertes : de cette manière, l'air passe à travers elles de façon continue, sans les faire vibrer, en causant une turbulence dans le conduit vocal. Les sons produits de cette façon sont dits "non voisés" et ne correspondent qu'à des consonnes telles que les fricatives (ch, f, s...) et certaines plosives (k, p, t...). Ces consonnes sont dites sourdes.

La parole est constituée d'une suite continue de sons voisés ou non, dont l'amplitude et le pitch, qui caractérisent la "source" de signal, varient en permanence, et quelquefois assez rapidement. Le signal émis par la source décrite ci-dessus est ensuite "filtré" par le conduit vocal constitué des cavités pharyngienne, buccale et nasale. Cette dernière n'est généralement pas prise en compte dans les synthétiseurs intégrés, pour simplifier leur réalisation.

On peut alors, en synthèse vocale, assimiler le conduit vocal à un tube de diamètre constant, dont les résonances principales sont représentées sur la figure 2.

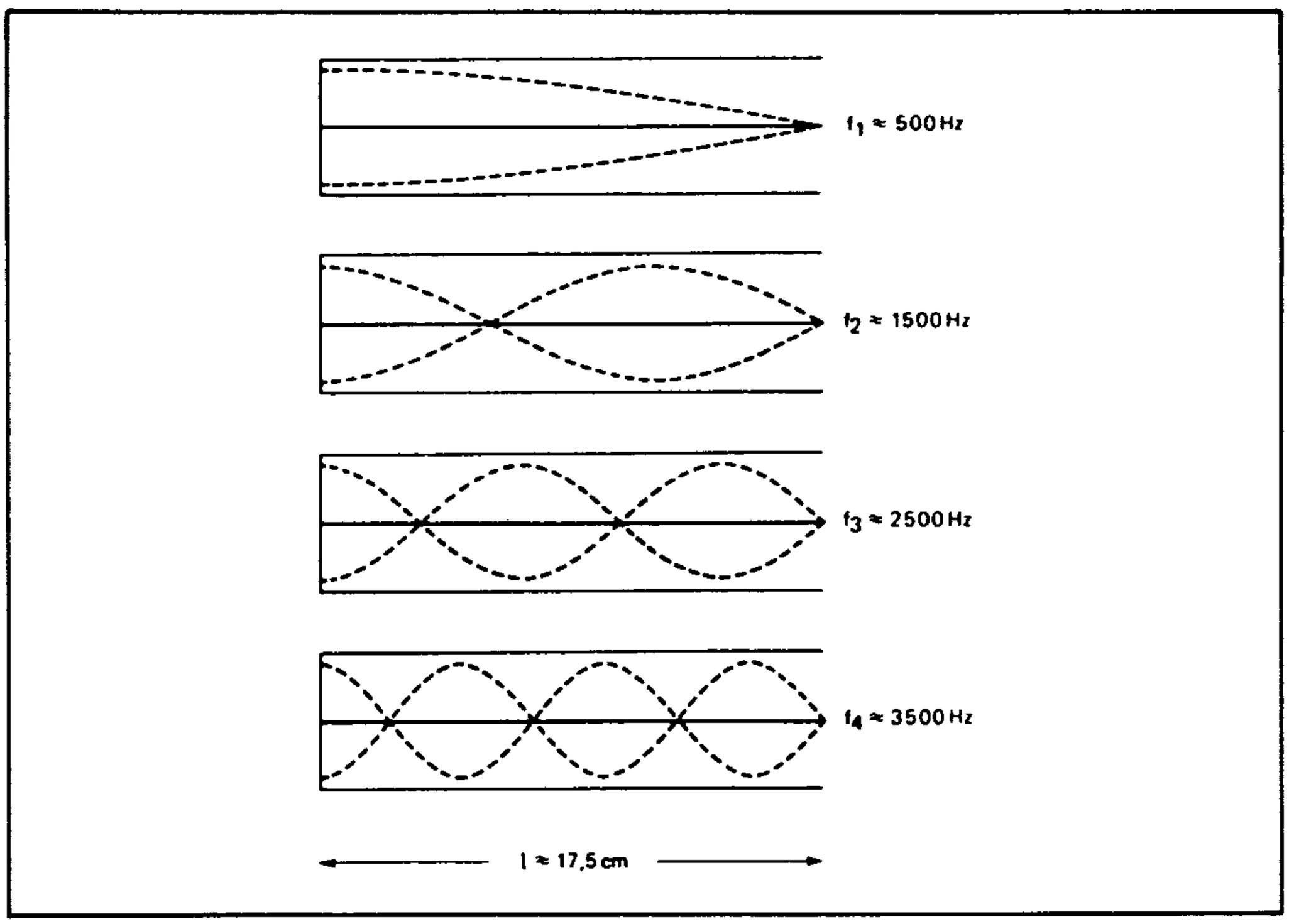

Figure 2 - Résonances d'un tube cylindrique

Ce tube est pratiquement fermé à l'extrémité constituée par les cordes vocales, et ouvert à l'autre par la bouche qui diffuse le son vers l'extérieur. La réponse en fréquence d'un tel tube est caractérisée par un nombre de résonances équidistantes dont les fréquences sont données par la relation :

$$f(N) = 340\,(2N - 1)/4L$$

où N = 1, 2, 3, 4 ... et L = longueur du conduit vocal (en mètres).

Ces fréquences de résonance sont appelées "formants" du conduit vocal. A l'intérieur de la bande de 0 à 4000 Hz, on trouve en général quatre formants pour une voix masculine et trois pour une voix féminine, en raison de la longueur plus réduite du conduit vocal chez la femme. La figure 3 illustre la position relative des formants dans les deux cas.

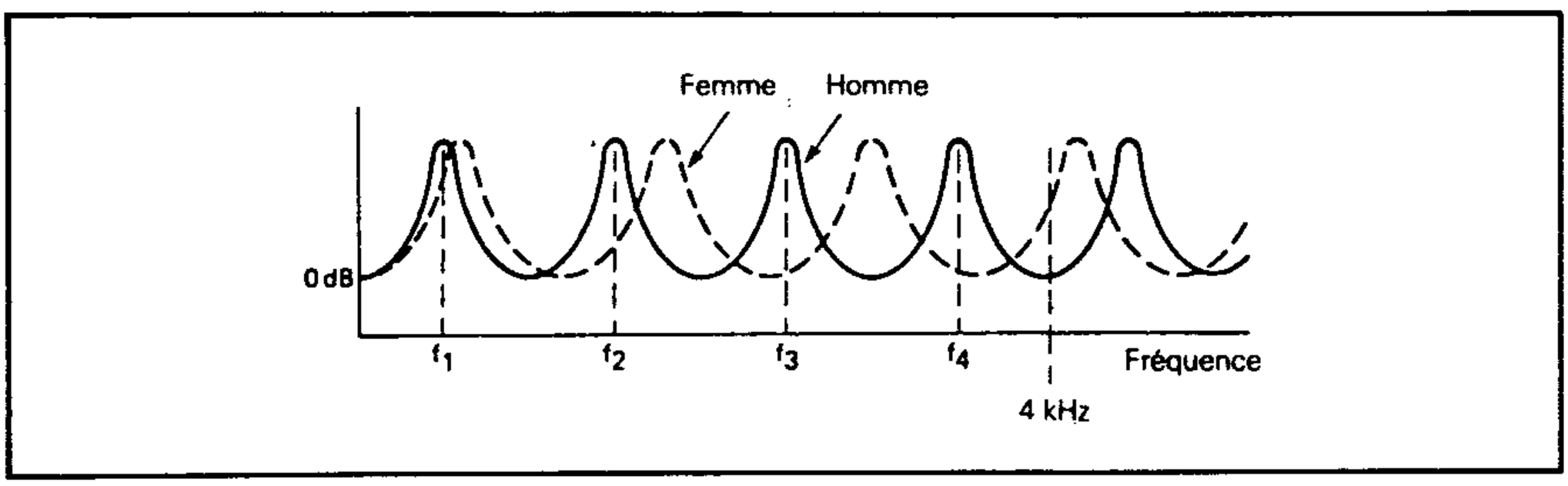

Figure 3 - Position des formants

Au cours de la parole, la forme du conduit vocal varie constamment : par exemple, pour le son "e", la cavité pharyngienne est grande alors que la cavité buccale est petite, ce qui a pour effet d'accroître la fréquence du formant n° 2. Lorsque l'on prononce un "a", la situation est inverse, ce qui réduit la séparation entre les formants n° 1 et 2. La figure 4 illustre ces deux situations.

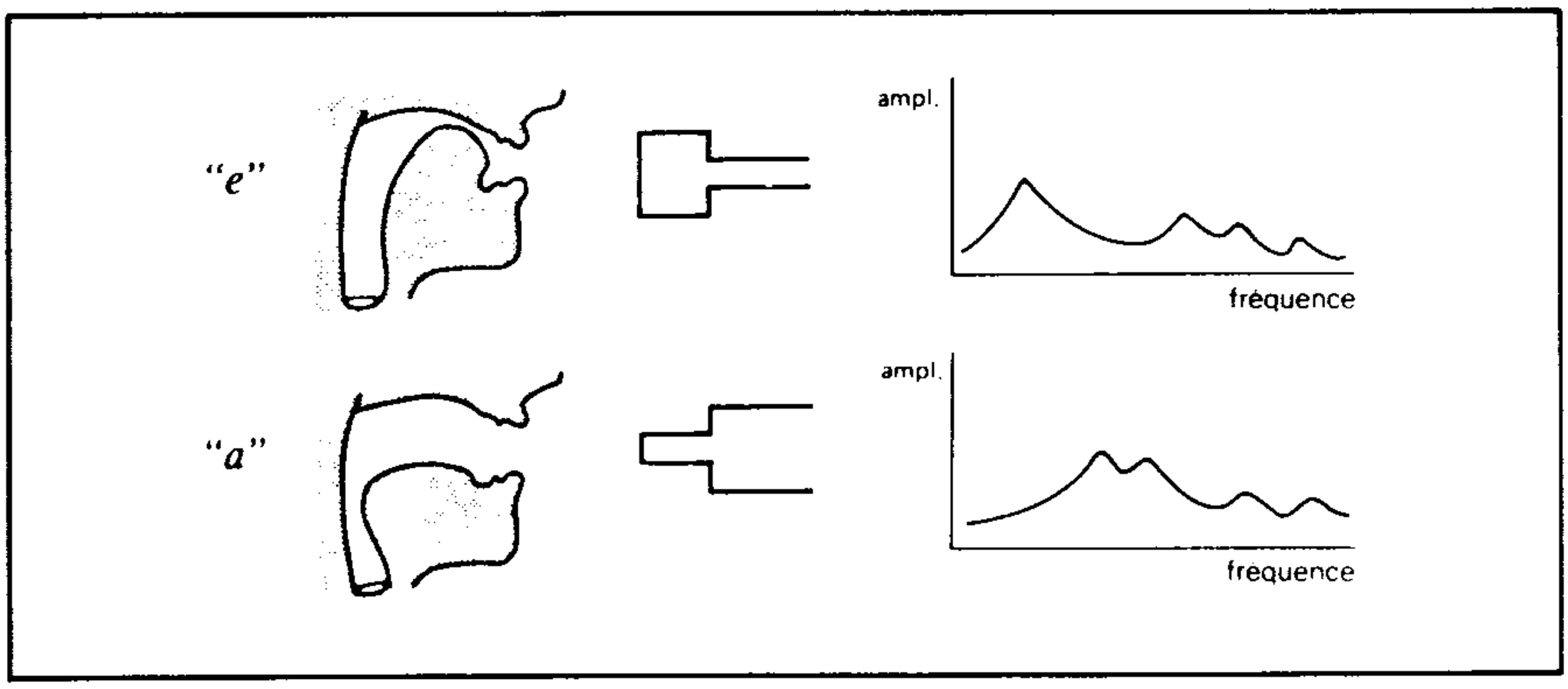

Figure 4 - Spectre des formants pour le "a" et le "e"

Chaque formant est en outre caractérisé par sa bande passante qui correspond à une résonance plus ou moins prononcée : plus la bande est étroite, plus la résonance est importante.

Ce sont les formants qui caractérisent réellement le son émis à un instant donné, et les trois premiers sont les plus importants pour l'intelligibilité du message.

Le fondamental (pitch), ainsi que, dans une certaine mesure, le voisement et l'amplitude, peuvent être considérés comme des informations "secondaires" pour la signification du message, essentiellement déterminée par l'évolution du conduit vocal et donc des formants. Les variations du pitch sont le facteur principal de l'intonation. Le non-voisement (en dehors des consonnes spécifiques) caractérise le chuchotement et l'accent tonique est déterminé par les variations instantanées de l'amplitude, dont de très rapides variations caractérisent également certains sons tels que les plosives.

De toutes les considérations précédentes, on peut déduire le schéma-bloc général d'un synthétiseur à formants (figure 5), qui n'est que la réalisation électronique simplifiée du modèle de départ.

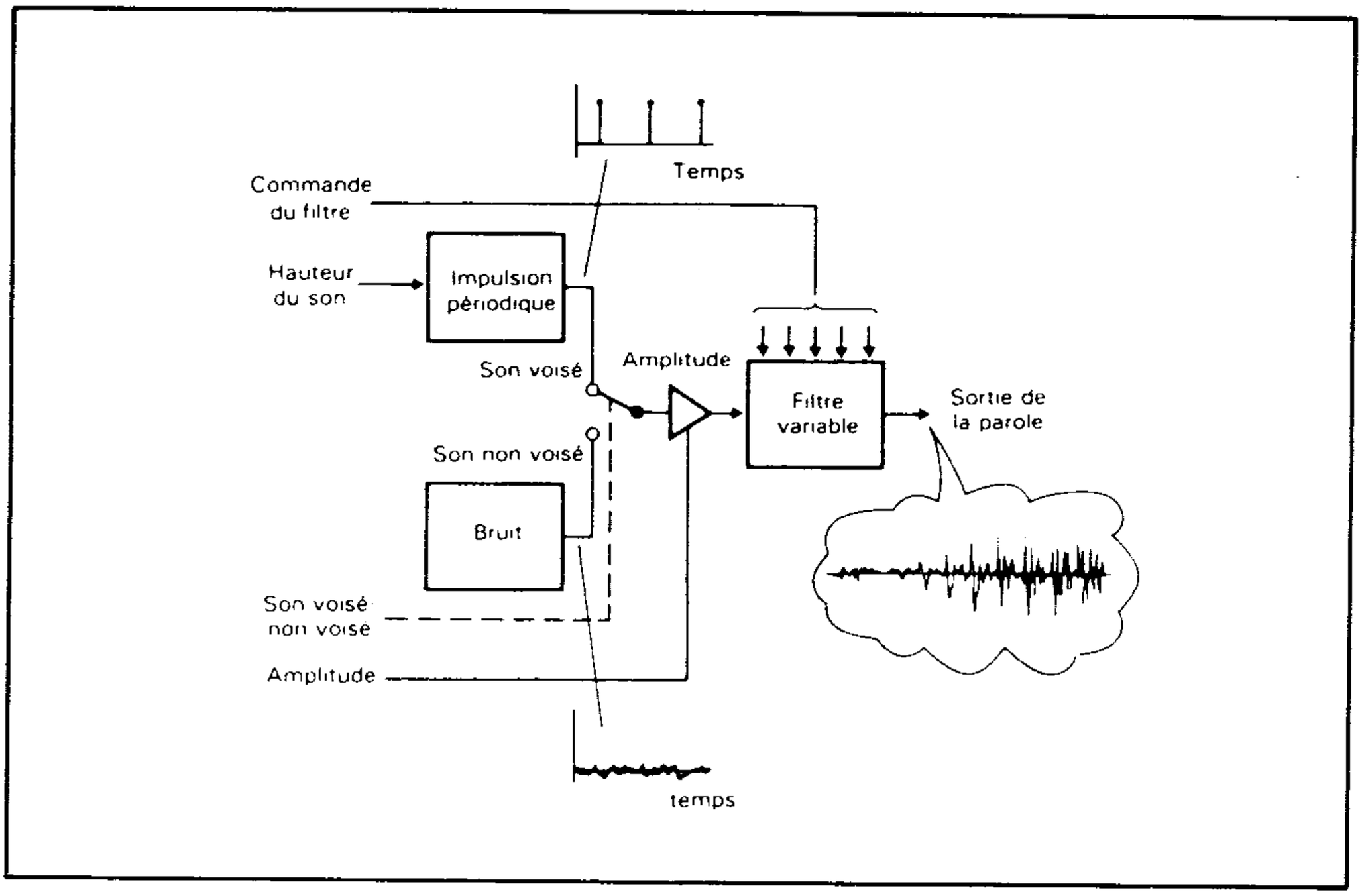

Figure 5 - Schéma-bloc d'un synthétiseur de parole

Le synthétiseur se compose donc d'une source de signal périodique, de fréquence programmable et égale au pitch de sons voisés. Pour la génération des sons non voisés, il dispose également d'une source de bruit blanc de 0 à 4000 Hz. Un commutateur permet de sélectionner l'une ou

l'autre source, et une commande de gain permet de déterminer l'amplitude du signal d'excitation. Enfin, un réseau de filtres programmables permet la simulation du conduit vocal. Chacun de ces filtres est programmable en fréquence et en bande passante et l'ensemble permet de modeler le spectre du signal conformément à la parole originale.

En résumé, le son émis par le synthétiseur à un instant donné est déterminé par l'ensemble des paramètres suivants:

— Fréquence fondamentale (pitch)	caractérisent la source d'excitation
— Son voisé ou non voisé	
— Amplitude (ou énergie)	
— Fréquence centrale des formants	caractérisent le conduit vocal
— Bande passante des formants	

L'obtention d'une reproduction fidèle de la parole à synthétiser nécessite une actualisation périodique de tous ces paramètres, considérés comme fixes dans une "fenêtre" dont la durée détermine la durée de trame sonore. Le conduit vocal humain étant un système "mécanique", ses variations sont relativement lentes, et l'actualisation des paramètres du synthétiseur peut être réalisée de manière satisfaisante avec une périodicité de l'ordre de 10 à 20 millisecondes.

Cette période (durée de trame) doit être suffisament longue de façon à contenir assez d'échantillons pour le calcul des paramètres lors de l'analyse du signal vocal, mais pas trop afin de pouvoir reproduire de manière satisfaisante ses variations les plus rapides.

Afin de ne pas provoquer de brusques changements lors du passage d'une trame à la suivante, une interpolation linéaire est souvent effectuée sur les paramètres entre deux trames consécutives. L'ensemble des fonctions du synthétiseur est entièrement réalisé en technique numérique, le signal de sortie étant restitué au moyen d'un convertisseur numérique/analogique (C.A.N.). Les messages à reproduire sont stockés dans une mémoire ROM ou RAM selon l'application.