HAMFU
HISTORY

KRIEG IM AETHER

Vorlesungen an der Eidgenössischen Technischen Hochschule in Zürich
imWintersemester 1965/1966
Leitung: Abteilung für Übermittlungstruppen, Oberstdivisionär E. Honegger

Notions sur les vocoders

Referent: E. Paris

Diese Vorlesung wurde durch die Stiftung HAMFU digitalisiert und als
PDF Dokument für www.hamfu.ch aufbereitet.

NOTIONS SUR LES VOCODERS

1. METHODES DE COMPRESSION DES SIGNAUX

En téléphonie classique, la parole est transmise par un signal électrique qui est la réplique du signal acoustique.

Si la parole devait être transmise sous la forme d'une transmission digitale - PCM par exemple - on estimerait aisément la vitesse à atteindre. En limitant la parole à un spectre de 3000 Hz et en se fixant une précision de l'ordre de 1 à 2 % sur l'amplitude du signal en fonction du temps, le nombre de bits par seconde sera donné par:

Fréquence d'échantillonnage 2 x 3000 = 6000/sec.

Nombre de niveaux 64, soit 6 bits

Soit, au total, 6000 x 6 = 36000 bits/sec.

Inversement, si l'on utilise la formule de Shannon,

$C = 2f_M \log_2 (1 + S/N)$ pour définir le canal de transmission capable de véhiculer 36000 bits, on constate qu'il est caractérisé par une bande passante de 3000 Hz avec un rapport signal à bruit en puissance de 30 dB.

$$C = 2 \times 3000 \log_2 (1 + 1000) \simeq 30\,000 \text{ bits/s}$$

Mais si l'on considère le taux d'information de la parole selon la manière dont elle est émise, l'on obtient un taux considérablement plus faible. Ainsi pour un langage utilisant 40 phonèmes dont nous supposons que la probabilité d'apparition est égale, le nombre de bits d'information est de 5,5 environ. En tenant compte du fait que l'appareil vocal n'est capable que de prononcer 10 phonèmes par seconde, la vitesse d'information est de 55 bits/sec.

Il y a donc une différence considérable entre le taux d'information nécessaire pour transmettre le signal vocal et celui qui est généré au niveau du mécanisme vocal.

Le signal vocal est donc particulièrement redondant et la difficulté réside dans la recherche des 55 bits de base parmi les 36 000 bits délivrés par l'échantillonnage temporel du signal vocal.

Cette redondance apparaît d'abord dans l'individualité de speaker qui, d'une manière continuelle, impose son identité et décrit son état émotionnel. Un autre aspect de la redondance se trouve dans les sons prolongés où une même forme d'onde est répétée plusieurs fois. Enfin, l'on peut dire aussi que le signal vocal n'utilise pas en permanence l'étendue maximum de son spectre.

Les méthodes de compression de la bande de fréquence du signal vocal tiendront compte de ces facteurs pour obtenir des gains sensibles sur la vitesse de transmission.

Parmi ces méthodes de compression, les principales sont:

- La compression dans le plan des temps ou dans le plan des fréquences.

  Le taux de compression à espérer, pour une qualité donnée de la parole, n'excède pas 2 à 3.

  En fait, ces méthodes n'utilisent pas à proprement parler les caractéristiques propres du signal vocal au lieu de sa création pour réduire le taux d'information. Ces méthodes sont applicables à n'importe quel type de signal.

- Méthode d'analyse et de synthèse continue.

  Dans ces méthodes, le signal vocal est décrit en fonction de ses paramètres propres. Elles exploitent le redondance du signal là où elle se crée. Les taux de compression peuvent dans ce cas atteindre 10 à 20. Les vocoders sont le meilleur exemple de cette technique.

- Méthode d'analyse et de synthèse discrète.

  De telles méthodes cherchent à décrire le signal vocal en le définissant à partir d'un groupe de sons élémentaires. Théoriquement elles devraient atteindre la limite possible de 55 bits.

  Le "pattern matching vocoder" est représentatif de cette technique.

2. LE MECANISME VOCAL

Le modèle le plus simple du mécanisme vocal peut se résumer dans le schéma donné en Figure 1.

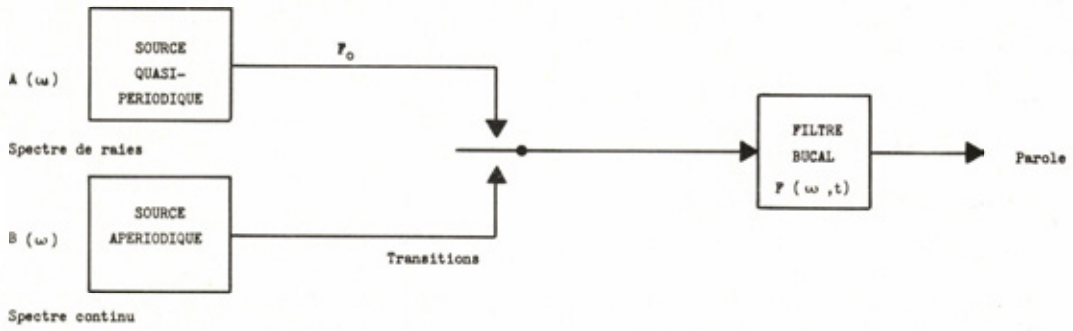

Fig. 1

La source A est constituée par l'ouverture et la fermeture des corde vocales. La durée 1/Fo entre deux impulsions successives est la période fondamentale de la voix (Pitch). Sa valeur se situe environ de 50 Hz à 400 Hz. Cette période est constamment variable pour donner à la voix un aspect esthétique (Pitch contour). Une voix monocorde conservera cette période constante.
Les voyelles sont créées à partir de la source A.

La source B est constituée par l'éjection de l'air de la bouche sans l'usage des cordes vocales. C'est le cas des sons du type s, ch, f. (fricatifs)

Les spectres d'énergie, constitués de raies pour la source A et continus pour la source B, sont modifiés par le filtre vocal F (ω , t) afin de créer le spectre final A (ω) x F (ω,t) ou B (ω) x F (ω,t) qui constituera la parole. (Figure 2).

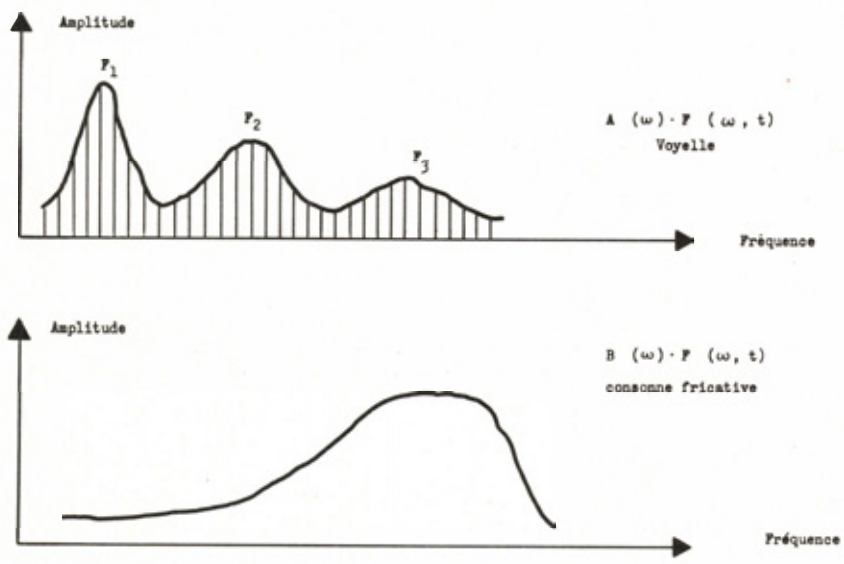

Fig. 2

Les maxima et minima du spectre de fréquences vocales sont respectivement appelés formants (F_1, F_2, F_3) et antiformants. Ils sont caractéristiques des sons qu'ils représentent.

On peut admettre que la configuration de ces spectres de fréquences ne change d'une manière significative que toutes les 40 ms. La description du mécanisme vocal comprendra donc:

- la mesure de la quasi périodicité de la source A
- les instants de transitions de la source A à la source B et vice versa.
- La description de l'enveloppe du spectre instantané de fréquences, soit F (ω,t).

3. LES PRINCIPAUX TYPES DE VOCODERS

3.1. Le "Spectrum Channel Vocoder"

La première description du Vocoder - Voice Coder - a été donnée par Homer Dudley (Bell Telephone Laboratories) en 1928. Le terme vocoder a été étendu aujourd'hui à l'ensemble des systèmes d'analyse et de synthèse de la parole.

Le Spectrum Channel Vocoder comporte deux parties essentielles, l'une relative à la pseudo fréquence fondamentale F_o et aux transitions A (ω) /B (ω), la seconde à l'enveloppe F (ω,t).

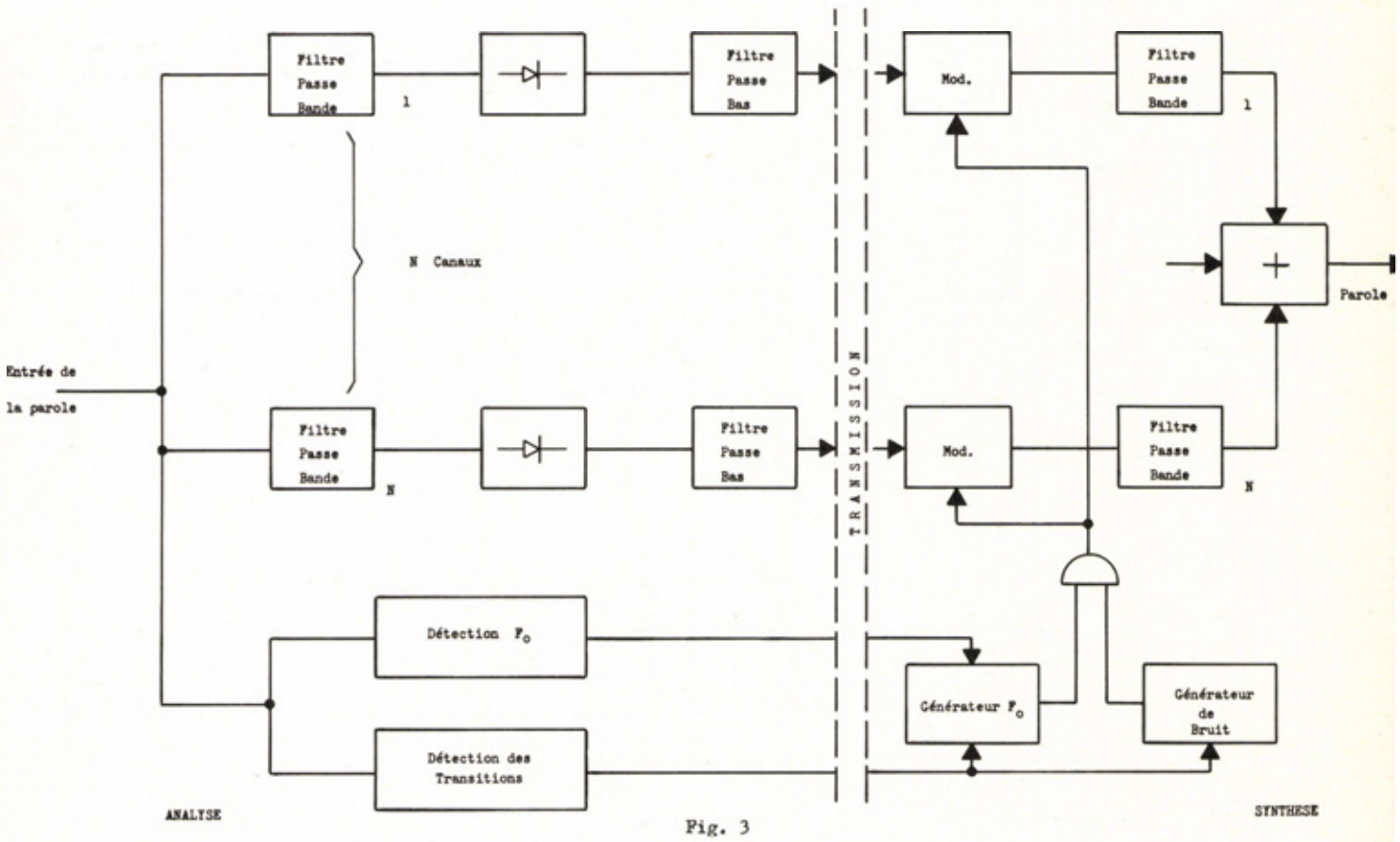

Fig. 3

Dans la partie supérieure gauche de la Figure 3, sont représentés les N canaux d'analyse. Chacun de ceux-ci comporte un filtre passe-bande, un redresseur et un filtre passe-bas. Le canal de rang j prélève une partie du spectre Δ fj, fonction du filtre passe-bande d'entrée, qui après redressement et intégration, décrira un point de l'enveloppe de F (ω,t), soit $|F_{\Delta fj}(\omega,t)|$

Les circuits de détection de la pseudo fréquence fondamentale et des transitions d'un mode à l'autre sont en général très complexes. Ils constituent la plus grande difficulté de l'analyse.

Parmi les principes les plus utilisés, laissant à part les méthodes de prélèvement directes comme le capteur du type laryngophone, on retrouve le filtrage plus ou moins complexe (filtre de poursuite, filtrage inverse), l'amplification non-linéaire du signal vocal pour concentrer l'énergie au voisinage de la fréquence fondamentale; l'analyse spectrale fine, et enfin, tout récemment, la méthode apparue sous le vocable "Cepstrum".

Les circuits de détection des transitions sont plus ou moins liés aux précédents, mais leur rôle capital est de décider de l'instant de la transition. De la bonne définition de cet instant dépendra en partie la qualité de la parole qui sera reconstituée.

La recomposition de la parole s'effectuera à l'aide d'un nombre identique de canaux comportant chacun un modulateur et un filtre passe-bande. En général, ce dernier est identique au filtre d'analyse (quant à sa bande passante).

Les filtres de synthèse peuvent être soumis à un régime impulsionnel dont l'amplitude est donnée par $|F_{\Delta f_j}(\omega, t)|$ et la récurrence par un générateur délivrant des impulsions à fréquence F_o ou suivant un régime aléatoire, selon les transitions observées.

Des aménagements peuvent être apportés au "channel vocoder" qui vient d'être décrit afin d'en améliorer l'intelligibilité et le ton naturel, mais ils n'en modifient pas le principe. Une des faiblesses de ce type de vocoder due à l'imperfection de la mesure de la fréquence fondamentale est le ton rugueux de la voix reconstituée.

D'autre part, la segmentation -ou le manque de gradation- des modes d'excitation des filtres sous la conduite des transitions nuit aussi au caractère naturel de la voix.

Lorsque l'analyseur et le synthétiseur doivent être reliés par un canal de transmission les méthodes classiques de modulation et multiplexage s'appliquent aisément. La transmission pourra être aussi bien digitale qu'analogique.

Le taux de compression d'information obtenu voisine 10 à 15. Les vitesse de transmission seront comprises entre 2000 bits/s et 5000 bits/s.

3.2. "Formant Tacking Vocoder"

Le "formant tacking vocoder" est un équipement de compression qui exploite d'avantage que le "channel vocoder" la redondance de la parole. Par exemple, durant la prononciation d'une voyelle on supposera qu'un ensemble de trois circuits résonants est suffisant pour décrire le filtre vocal $F(\omega, t)$. Par conséquent, six paramètres seront transmis qui sont la position des trois formant (F_1, F_2, F_3), au long de l'axe des fréquences et leurs amplitudes respectives A_1, A_2, A_3.

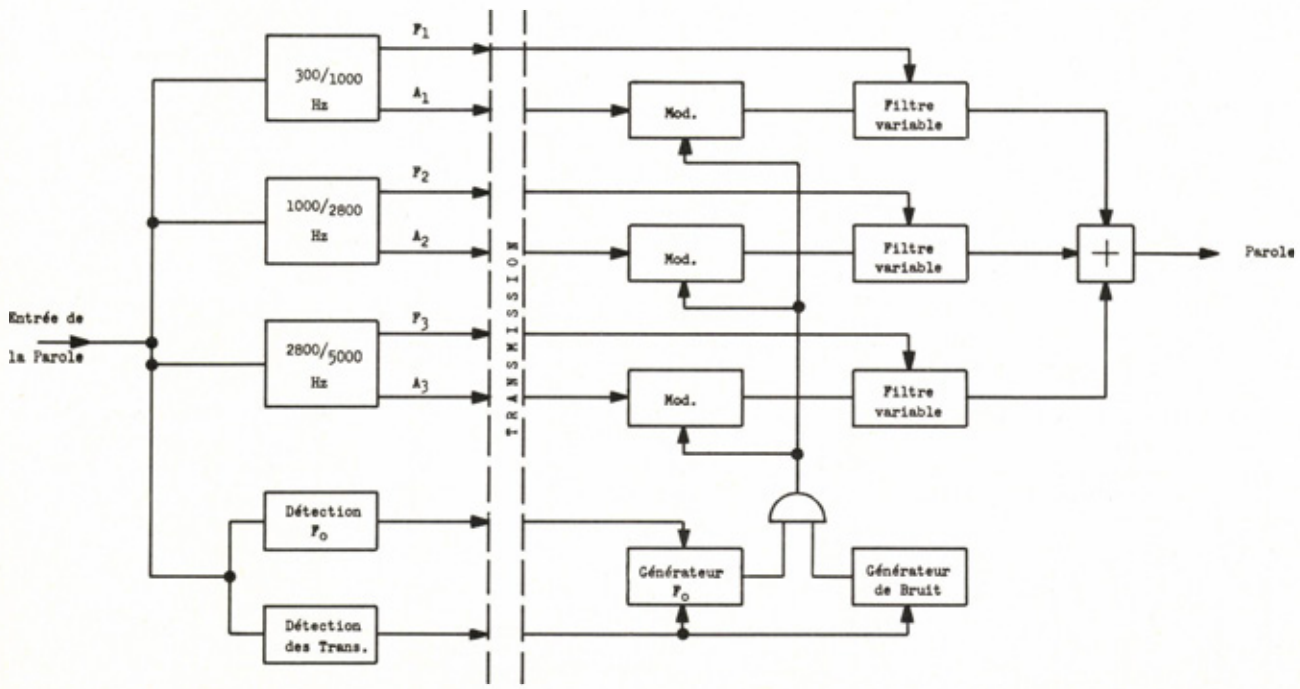

Fig. 4

La Figure 4 décrit un type de "formant tracking vocoder" dans lequel l'arrangement est fait en parallèle. Il existe d'autres équipements où l'arrangement est du type série.

Les circuits détecteurs de F_0 et des transitions sont identiques à ceux du "channel vocoder".

Des vitesses de transmission de 1200 bits/sec. peuvent être atteintes avec de type de vocoder, mais la qualité est inférieure à celle du channel vocoder. La raison est que bien des sons ne peuvent se représenter simplement par trois résonances, c'est le cas des semi-voyelles (r, l,) et des nasales (n, m,).

Pour pallier ce défaut, des améliorations peuvent être apportées, notamment en augmentant le nombre de paramètres à transmettre et en utilisant des circuits spéciaux pour les consonnes nasales, mais la vitesse de transmission sera augmentée.

3.3. Le "voice-excited vocoder" (V.E.V.)

Le manque de naturel des deux vocoders précédents a son origine dans l'imperfection des circuits de détermination de F_0 et des transitions.

Le siège de ces difficultés réside dans le fait que F_0 peut couvrir une gamme très large dans des conditions très variables d'amplitude. De plus, les phénomènes de réverbération, de bruit ambiant sont une source considérable d'erreurs.

Aussi a-t-il été dépensé beaucoup d'imagination pour contourner ces difficultés. Une solution élégante a été proposée par les Laboratoires de la Bell Telephone (Schroeder et David) sous le nom de "voice-excited vocoder". Le principe consiste à tronquer le signal vocal en deux spectres. Le premier (300 - 900 Hz) -la bande de base- sera transmis sans compression tandis que le second (1000-3500) sera analysé, par exemple, avec l'une des méthodes décrites en 2.1. ou 2.2.

Au niveau de la réception, la bande de base est traitée suivant un processus non-linéaire, afin de la transformer en un spectre plat, $P(\omega) = cte$, sur l'étendue du spectre vocal (Fig. 5).

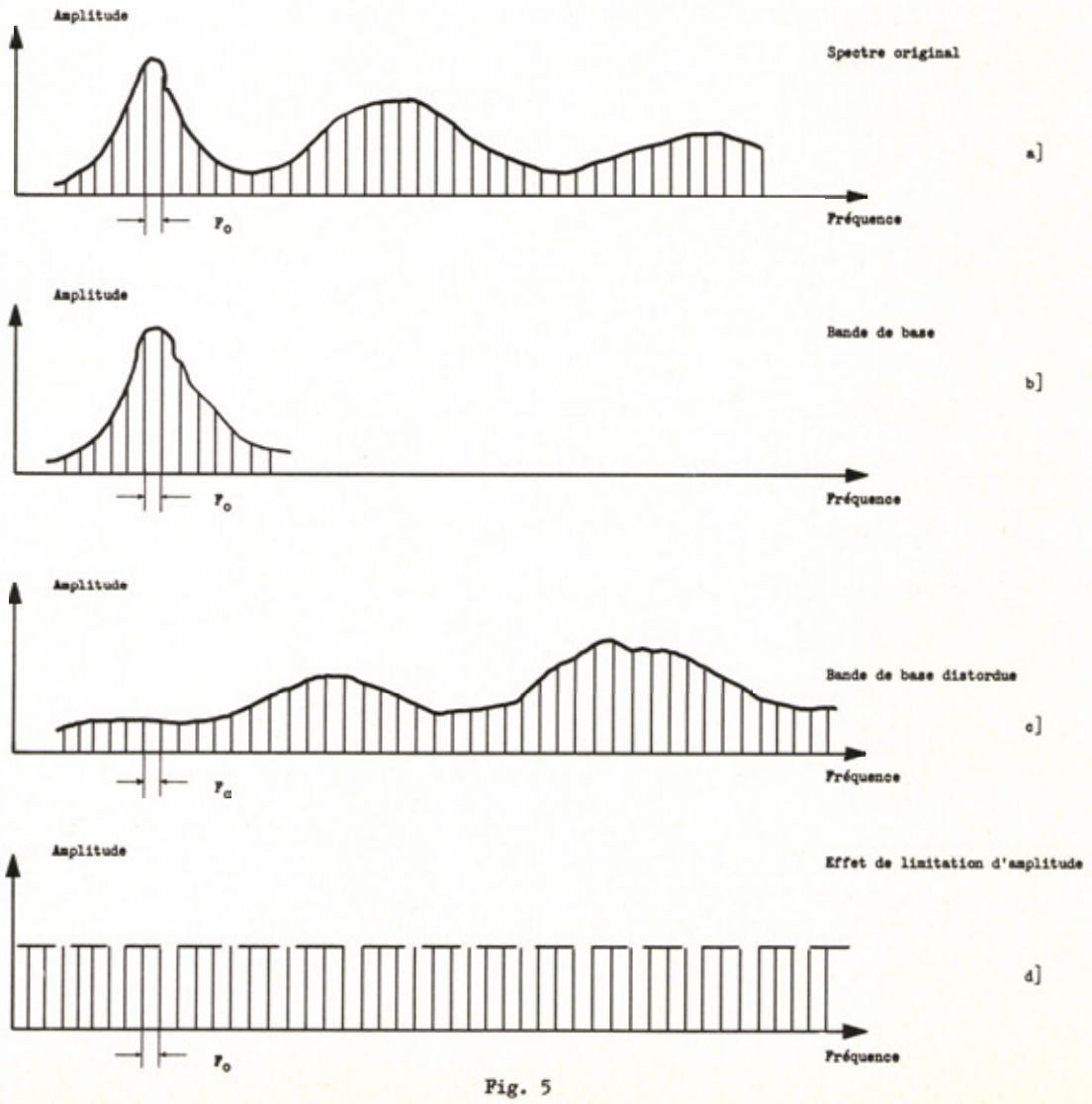

Fig. 5

L'effet de limitation dans le V.E.V. permettra de reproduire automatiquement la transition des modes d'excitation. Si la pseudo-fréquence fondamentale F_o disparaît le signal devient ou apériodique -spectre continu- ou zéro; les transitions seront donc beaucoup plus graduelles que dans les vocoders traditionnels.

La figure 6 montre un diagramme du V.E.V.

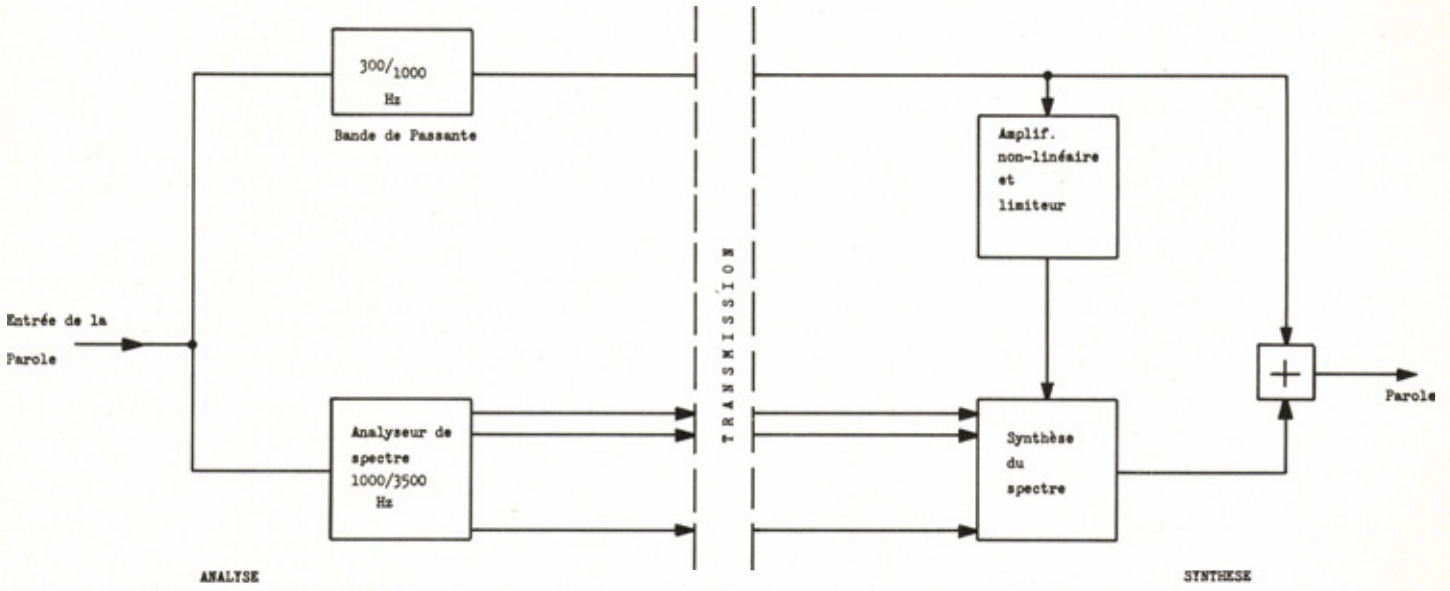

Fig. 6

Les vitesses de transmission nécessitées par le V.E.V. sont voisines de 10 000 bits/s. On le voit aisément dans l'exemple suivant:

- bande de base: 300 - 900 Hz soit 7 200 bits/s pour une quantification à 16 niveaux.

- spectre supérieur: pour un ensemble de 15 filtres suivit de filtres passe-bas ayant une fréquence maximum transmise de 25 Hz, et pour une quantification à 8 niveaux, on trouve 2 250 bits/s soit un total de 9 450 bits/sec.

3.4. Différents types de vocoders

3.4.1. Auto-corrélation vocoder (Fig. 7)

Le "channel vocoder" montre que l'intelligibilité du signal vocal f(t) est transportée par l'amplitude du spectre instantané de fréquence $F(\omega, t)$.

La description de ce spectre est obtenue par un ensemble de filtres passe-bande. Mais toute description équivalente de la fonction désirée sera acceptable. C'est ainsi que la transformée de Fourier relie simplement le spectre d'énergie d'un signal $F(\omega, t)$ 2 à sa fonction d'autocorrélation $\varphi(\tau, t)$. On pourra donc à partir d'une analyse du signal dans le plan des temps décrire l'amplitude du spectre instantané du signal par sa fonction d'autocorrélation instantanée. Toutefois, une légère transformation sera nécessaire pour que la fonction d'auto-corrélation décrive $F(\omega, t)$ et non $F(\omega, t)$ 2.

Dans ce type de vocoder, la détermination de la fréquence fondamentale et des transitions, s'opère comme dans un "channel vocoder".

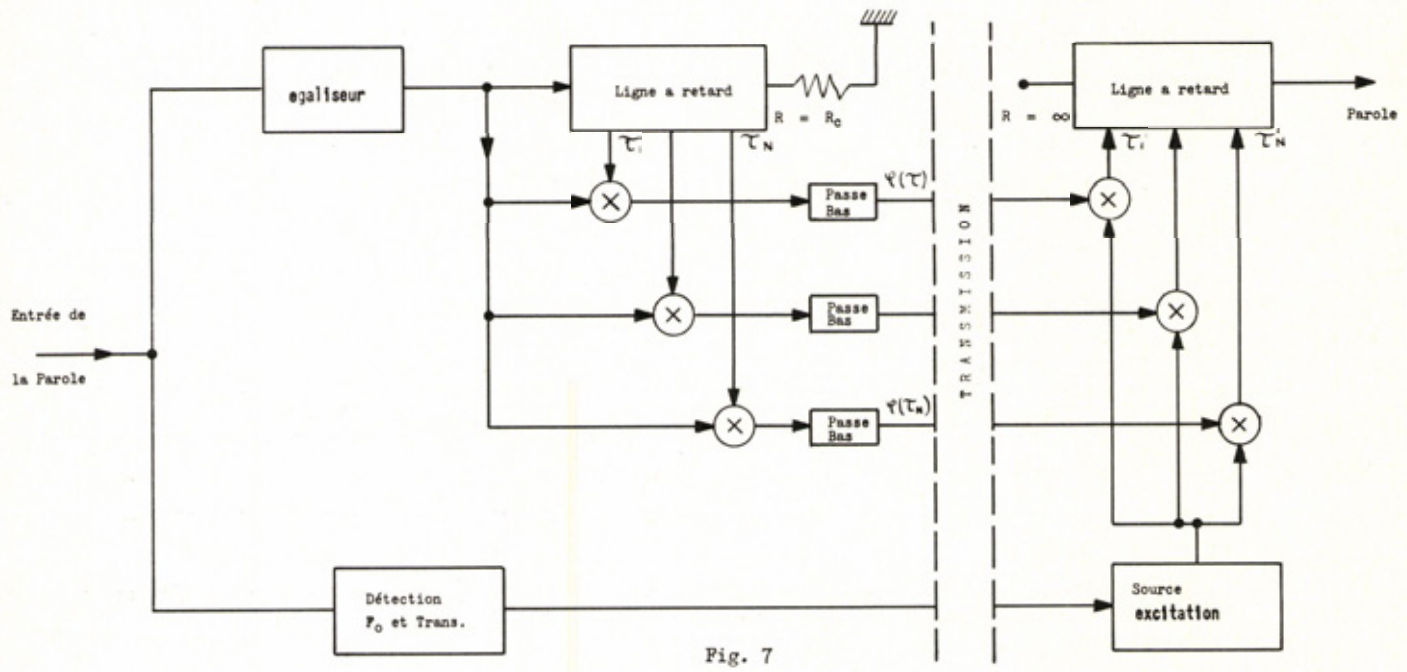

Fig. 7

Des filtres passe-bas limiteront le spectre de $\Psi(\tau_i)$ à 25 Hz environ. Le nombre N de prises sur la ligne à retard sera déterminé en fonction du théorème d'échantillonnage. Si par exemple le signal vocal est analysé par tranche de Θ = 5 ms, et si l'on admet que la fréquence maximum f_c transportée par le signal vocal est de 3 000 Hz:

$$N = 2\Theta f_c = 2 \times 3\,000 \times 5 \quad 10^{-3} = 30$$

Pour restituer le signal, une ligne à retard identique est utilisée: elle opérera la somme des produits des fonctions $\Psi(\tau_i)$ par la source d'excitation. Afin de rendre à la fonction d'autocorrélation sa symétrie temporelle, la ligne à retard sera ouverte à une extrémité.

Le taux d'information obtenu après analyse est à peu près équivalent à celui d'un "channel vocoder".

3.4.2. Le "Pattern Matching Vocoder"

Ce vocoder procède à une analyse discrète du spectre vocal instantané.

Le spectre vocal est comparé à un ensemble de N spectres prédéterminés parmi lesquels on recherche celui qui délivre le taux de corrélation maximum. Un code est associé à chaque spectre prédéterminé. (Fig. 8)

La détection de F_o et des transitions est identique à celle du channel vocoder.

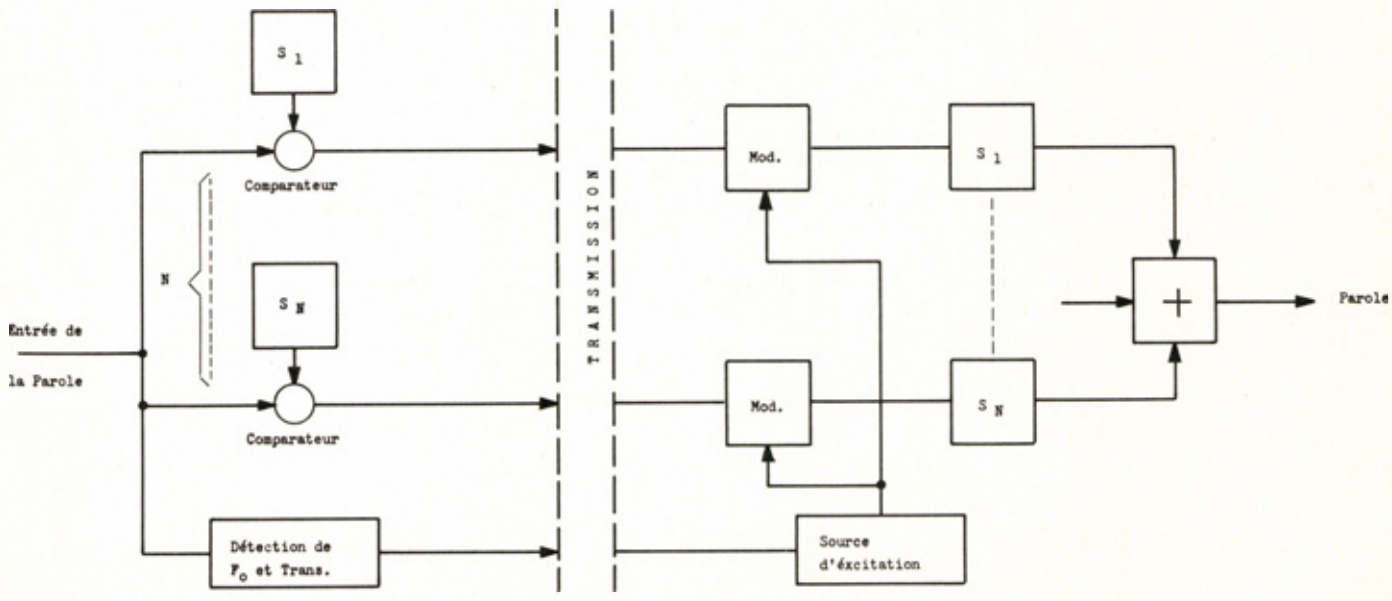

Fig. 8

Une des difficultés inhérentes à ce type de vocoder sont les transitions d'un spectre à un autre au cours de la recomposition. Des filtres, non représentés ici, sont nécessaires pour assurer des passages graduels dont se satisfasse l'oreille.

Les taux d'information obtenus sont voisins de quelques centaines de bits par seconde. Le taux de compression sera donc de l'ordre de 50, ce qui est considérable.

4. EVALUATION DE LA QUALITE DES VOCODERS

4.1. Intelligibilité

Il existe plusieurs méthodes pour évaluer la qualité des vocoders, mais, en principe, elles ne diffèrent que par la nature des signaux vocaux utilisés. Ces signaux pourront être ou des messages, ou des mots isolés, ou encore des sons de base dépourvus de contenu sémantique. Le choix dépendra de la finesse de l'évaluation que l'on voudra obtenir.

La mesure de l'intelligibilité sera le pourcentage des signaux reconnus parmi l'ensemble des signaux proposés.

Il existe des standards qui définissent les niveaux sonores à utiliser, rapport signal à bruit d'ambiance ainsi que les conditions d'entraînement des auditeurs.

4.2. Préférence

L'évaluation de la préférence est beaucoup plus subjective. Elle consiste à rechercher à partir de quel niveau de dégradation le signal vocal original n'est plus préféré au signal vocal délivré par le vocoder.

La dégradation du signal vocal original peut être obtenue soit par addition ou multiplication de bruit blanc filtré.

Cette méthode permet de chiffrer (en db par exemple) une propriété purement subjective. Elle permet entre autre d'établir des comparaisons entre différentes voix synthétiques.

5. CONCLUSION

En matière de conclusion, notre attention se porte vers quelques points précis.

- La réduction du taux d'information a des prolongements évidents en matière d'économie. Si l'information transportée par le signal vocal peut être concentrée en une bande de fréquences très étroite, la capacité des canaux de transmission se trouve grandement accrue.

  Ceci devrait représenter dans le futur un allégement pour les capitaux d'investissement en matière de systèmes de transmission. Des améliorations substantielles devraient pouvoir aussi être apportées aux équipements mobiles de transmission.

- La transmission digitale des signaux délivrés par les vocoders mérite d'être comparée à la transmission analogique classique. Un signal digital est incomparablement moins sensible aux imperfections de la transmission qu'un signal analogique. Le codage et les circuits de détection et de correction d'erreurs sont des outils puissants pour se prémunir contre les aléas de la transmission. On imagine aussi aisément la simplicité d'un répéteur téléphonique chargé de retransmettre des signaux digitaux.

- La représentation digitale du signal vocal se prête admirablement aux procédés de cryptage. Les possibilités offertes aux spécialistes sont infiniment supérieures et plus variées que dans le cas du signal analogique.

  Enfin, les paramètres de la voix, dans leur représentation digitale, peuvent être mis sous le contrôle des calculateurs. On devine aisément les ressources de cette association qui commencent à poindre aujourd'hui et dont le futur ne manquera pas d'être étoffé.

Annexe bibliographique

- Livres

Experiments in Hearing	G. Von Bekesy	(Mc Graw-Hill)
Acoustical Measurements	L.L. Beranek	(J. Wiley)
Das ohr ald nachrichtchenemp-pfänger	R. Feldtkeller E. Zwicker	(Hirzel, Stuttgart)
Speech analysis synthesis and perception	J.L. Flanagan	(Springer, Berlin)
Speech and hearing	H. Fletcher	(D.V. Nostrand)
Acoustical Engineering	H.F. Olson	(D.V. Nostrand)
Vibrations en Régime Transitoire	L. Pimonow	(Dunod, Paris)
Visible Speech	R. Potter	(D.V. Nostrand)

Revues spécialisées

- The journal of Acoustical Society of America
- The journal of the Audio Engineering Society
- I E E E Transactions on Audio
- Proceedings of the International Congress on Acoustics
- Acoustica
- Frequenz
- N T Z