Comment ça marche ?

Les transmissions radio-numériques

16 – Le codage de la parole (2)

Par le radio-club F6KRK

Après avoir vu en première partie le codage de la parole à haut débit, nous allons poursuivre avec les codages à moyens et bas débits. Bien sûr, il ne s'agit pas d'exiger du lecteur une compétence en traitement du signal. Cet article n'a d'autre ambition que de donner une idée générale des processus employés.

Le codage CELP (moyens débits)

N-B : *La réduction du débit entraîne la nécessité de découper le flot de données en paquets : les trames. Nous sommes alors en temps différé (retard du à la constitution d'un historique) et aucune erreur n'est tolérée.*
Pour ces débits réduits, les techniques de codage de la forme d'onde que nous avons vus ne donnent pas de bons résultats. Les codeurs doivent éliminer les informations sans pertinence pour la perception. Les vocodeurs utilisent certaines caractéristiques de la perception et de la production de la parole, aussi sont-ils généralement très peu efficaces pour les signaux autres que la parole comme du bruit ambiant ou des signaux DTMF.
La solution la plus employée utilise le codage **CELP** (**C**ode **E**xcited **L**inear **P**rediction). Il est très efficace pour les débits moyens de 4,8 à 16 kbps, comme en témoignent les nombreuses normes qui l'utilisent. La figure 1 représente le principe de ce codage.

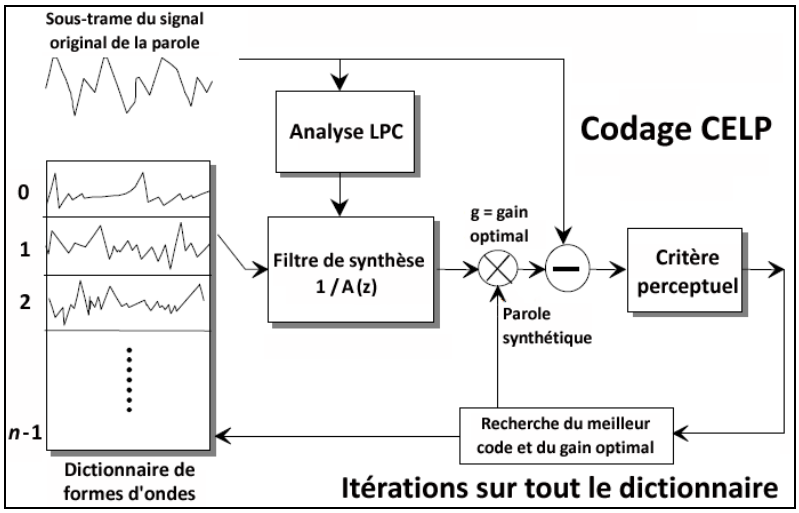

Figure 1 : **Principe du codage CELP**

Dans chaque trame, une analyse spectrale par prédiction linéaire (LPC) détermine le filtre de synthèse *1/A(z)*. On découpe chaque trame en sous-trames plus courtes (durée typique 5 ms) sur lesquelles on effectue une quantification vectorielle du signal par une technique d'analyse

par synthèse. On compare à l'aide d'un critère dit « perceptuel » [(1)], de type moindres carrés pondérés, le signal de parole original avec tous les signaux synthétiques possibles obtenus après quantification vectorielle. Ces signaux synthétiques sont générés en filtrant par le filtre de synthèse, un signal d'excitation choisi dans un dictionnaire de séquences d'excitation (on ajoute parfois la sortie de plusieurs dictionnaires) et en ajustant le signal résultant par le gain optimal. Le codeur transmet le ou les index des segments qui minimisent le critère ainsi que le ou les gains associés, les paramètres spectraux et le pitch fractionnaire (fréquence fondamentale *Fo*). Le critère perceptuel prend en compte la propriété de masquage du bruit de quantification par les formants en pondérant plus fortement l'erreur de quantification dans les zones de faible amplitude du spectre et plus faiblement dans les zones de formants. Cette pondération s'effectue en filtrant le signal d'erreur par un filtre de type $A(z)/A(z/\boldsymbol{k})$ où $\boldsymbol{k}$ est compris entre 0 et 1 (typiquement $\boldsymbol{k} = 0{,}85$). Les dictionnaires utilisés sont appelés stochastiques ou adaptatifs selon qu'ils contiennent des séquences fixes de bruit ou bien les séquences d'excitation de trames précédentes. Le dictionnaire adaptatif permet de prendre en compte la redondance introduite par la quasi-périodicité des sons voisés.
Pour les sons voisés, le signal synthétique présente des harmoniques depuis F_0 jusqu'à $Fe/2$ même si le signal original n'a plus d'harmoniques au-delà d'une fréquence *Fmax*. On parle dans ce cas d'artéfact tonal.
Pour le décodage, à partir des paramètres transmis, on utilise les numéros de forme du même dictionnaire avec leur gain puis on les additionne.

Les VOCODEURS (bas débits)

Les vocodeurs n'étant pas employés dans les transmissions radioamateurs, nous ne ferons qu'un survol général du sujet. Par ailleurs la complexité des traitements mathématiques utilisés les réserve aux spécialistes.
La qualité subjective des codeurs CELP décroît rapidement lorsque le débit descend en dessous de 4 kbps. En effet, le codage CELP effectue essentiellement une quantification vectorielle de la forme d'onde et pour un débit trop faible il n'est pas possible de coder cette forme précisément. Pour des faibles débits, on a recours à des vocodeurs (VOice CODER). Dans les vocodeurs classiques, vocodeurs à canaux, vocodeurs à formants, ou vocodeurs LPC, les différentes trames de signal sont classées en trames voisées (V) et trames non voisées (NV). Ces vocodeurs utilisent un modèle "source-filtre". La synthèse du signal décodé utilise un signal d'excitation reconstruit formé d'un bruit blanc pour les trames non-voisées et d'un train périodique d'impulsions à la fréquence F_0 pour les trames voisées. La figure 2 représente le synthétiseur d'un vocodeur à 2 états d'excitation.

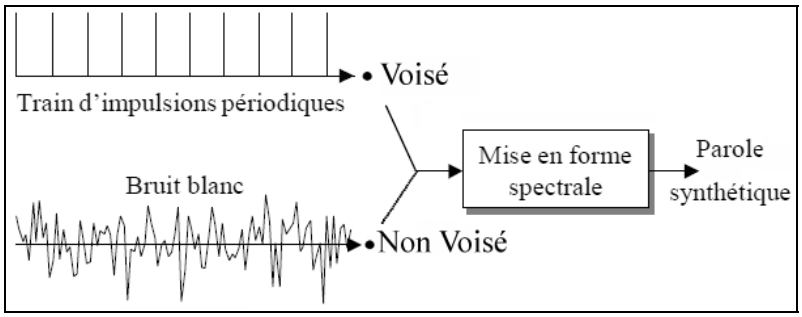

Figure 2 : **Synthèse à deux états d'excitation**

Vocodeurs à canaux

Dans les vocodeurs à canaux introduits par Dudley en 1939, le codeur évalue l'énergie, le voisement, F_0, et les puissances relatives du signal dans un ensemble de bandes de fréquences

adjacentes (de l'ordre de 10 bandes). Le décodeur génère la parole synthétique en passant le signal d'excitation dans un banc de filtres passe-bande dont les sorties sont pondérées par les puissances relatives du signal original dans ces différentes bandes. Les sorties des filtres sont ensuite ajoutées et cette somme est mise à l'échelle en fonction de l'énergie de la trame originale. Ces codeurs ont été utilisés jusqu'à des débits de 400 bps.

Vocodeurs à formants

Dans les vocodeurs à formants, le codeur détermine la position, l'amplitude et la largeur de bande des 3 premiers formants, ainsi que l'énergie de la trame, le voisement et *F0*. Les formants sont des bandes de fréquences dans lesquelles l'énergie de la parole se répartit. Elles dépendent de la constitution physique de l'appareil phonique du locuteur. Voir un exemple sur la figure 3 [(2)].

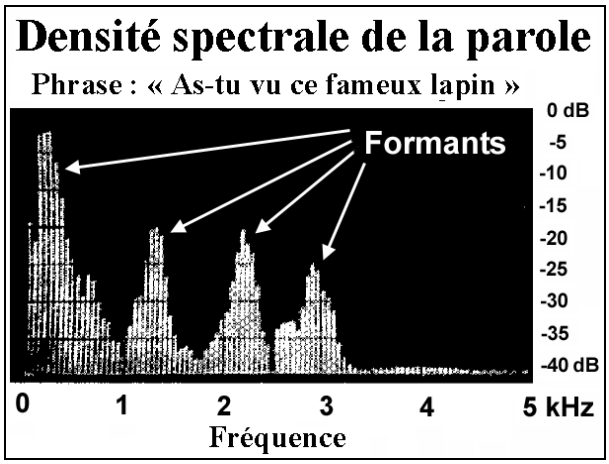

Figure 3 : **Aanalyse statistique des formants pour une phrase type et la voix de l'auteur**

Au décodage, l'excitation synthétique est filtrée par 3 filtres accordés sur les trois premiers formants. Le signal résultant est mis à l'échelle en fonction de l'énergie de la trame. On obtient avec cette technique un signal intelligible pour des débits de 1200 bps, mais la détermination des formants est une tâche difficile et peu fiable.

Vocodeurs à prédiction linéaire LPC

Succinctement : Dans les vocodeurs à prédiction linéaire LPC (Linear Predictive Coding), l'enveloppe spectrale du signal de parole est modélisée par l'amplitude de la fonction de transfert d'un filtre tout pôle ***1/A(z)***. Les coefficients a_i du filtre sont obtenus par prédiction linéaire à partir d'une combinaison linéaire des échantillons précédents. L'enveloppe spectrale est très sensible à la quantification des coefficients a_i. Le nombre de coefficients a_i est compris entre 8 et 16 pour une fréquence de 8 kHz, de façon à ce que la fonction de transfert du filtre présente un nombre suffisant de résonances pour modéliser correctement les 3 à 5 premiers formants. En plus des coefficients déduits des coefficients LPC, le codeur transmet l'énergie, le voisement et la fréquence fondamentale de la trame.
Le décodeur génère le signal synthétique en filtrant l'excitation reconstruite par le filtre de synthèse *1/A(z)* et en mettant à l'échelle la sortie en fonction de l'énergie de la trame.
Les codeurs LPC à 2 états ont été développés pour des débits d'environ 2400 bps. Des débits de 600 à 800 bps ont été atteints en appliquant une quantification vectorielle aux coefficients spectraux.
Là aussi nous nous sommes limités aux principes généraux. A partir de ceux-ci il y a une infinité de variantes qui ne dépendent bien souvent que d'un choix différent de paramètres [(3)].

Le MOS

La voix sortant des vocodeurs est une voix artificielle s'efforçant de reproduire le mieux possible la voix du locuteur qui comporte beaucoup d'informations subjectives : timbre, émotion, etc. En conséquence il n'est pas possible de noter la qualité d'un vocodeur à partir de mesures physiques. On a alors recours à un jury de personnes qui notent leurs "impressions" pour des utilisations particulières (radiodiffusion, téléphone, télécom, synthétique). Le résultat constitue le MOS (**M**ean **O**pinion **S**core) qui est une note qui varie de 0 à 5. Un MOS de 3,5 est considéré comme tout à fait acceptable (GSM-Fr). Noter qu'il varie selon la langue [(4)].

Dans le prochain "Comment ça marche ?" nous aborderons le codage des images fixes.

La Rubrique "Comment ça marche" est une activité collective du radio-club F6KRK (http://www.f6krk.org). Pour une correspondance technique concernant cette rubrique : "f5nb@orange.fr".

Bibliographie

Pour cette deuxième partie, nous nous sommes principalement servi de cette référence :
- Rapport d'étude "Codage de la parole à bas et très bas débit" par :
  - Geneviève Baudoin, ESIEE, Département Signaux et Télécommunications.
  - J. Cernoky, Université Technique de BRNO (République Tchèque).
  - P. Gournay, Thomson-CSF Communications (GENNEVILLIERS).
  - G. Chollet, CNRS URA-820, ENST-TSI (PARIS)

Notes :

1) *On utilise le néologisme « perceptuel » pour indiquer un critère ou un filtre essayant de tenir compte de la perception auditive qui n'est pas linéaire.*
2) *On a ici une voix d'homme pour laquelle un filtre BLU passe aisément les trois premiers formants. Pour une voix féminine plus haut perchée, le filtre BLU ne passe bien que deux formants, ce qui explique la plus grande difficulté à décoder les voix d'YL dans le cas d'un faible rapport S/B.*
3) *Citons en quelques uns : "RELP", "MLPC", "CELP", "VSELP"...*
4) *Un vocodeur avec un MOS de 3,9 en français peut très bien être noté 1,9 en chinois mandarin.*