Un système pour l'amélioration des communications en plongée profonde.

J. Crestel, M. Guitton

Laboratoire d'Analyse des Systèmes de Traitement de l'Information.
BP. 150 Avenue de la Résistance 22302 Lannion

Résumé

Un algorithme d'amélioration de l'intelligibilité de la parole "hyperbare" et l'architecture d'un système de type MIMD approprié à son exécution en temps réel sont proposés. L'algorithme exploite la décomposition de l'altération de la fonction de transfert du conduit vocal (selon l'axe des fréquences) en deux composantes -l'une linéaire, l'autre non linéaire- et s'appuie en particulier sur une modélisation du conduit vocal par un filtre tout-pôle caractérisable par les propriétés de la prédiction linéaire.

Abstract

This paper introduces an algorithm which improves the intelligibility of speech in helium environment as well as a system like MIMD well-suited for its execution in a real-time processing mode. The algorithm uses the division of the modification of the vocal tract transfer function (along the frequency axis) into two components -one is linear, the other is non linear- and is based on the application of the properties of linear prediction to an all-pole model of the vocal tract.

1- Nature de l'altération de la parole.

Il est connu que l'usage de mélanges respiratoires de type "héliox" ou "hydrox" rend la parole des plongeurs quasiment inintelligible. L'analyse du signal de parole met en évidence plusieurs types d'altérations responsables de la dégradation de l'intelligibilité [1] :

- les formants sont déplacés, non linéairement, vers les fréquences élevées .
- l'intensité des sons non voisés est atténuée, relativement à celle des sons voisés .
- une augmentation de la fréquence fondamentale est systématiquement observée. A ces phénomènes à priori justifiables par la théorie de production de la parole s'ajoutent d'autres facteurs de dégradation propres au contexte de la plongée -bruits, effets du masque de plongée, désordres comportementaux du plongeur qui peuvent apparaitre en plongée profonde et perturber l'élocution.

Le déplacement des formants, facteur essentiel de la dégradation de l'intelligibilité de la parole, induit une modification de la distribution spectrale de l'énergie du signal de parole: un traitement opérant une correction dans le domaine fréquentiel semble donc mieux approprié qu'un traitement direct dans le domaine temporel, ce que confirme la qualité de certaines simulations ou réalisations de décodeurs prototypes [2,3,4]. La solution préconisée s'appuie sur une modélisation de l'appareil phonatoire et sur une formulation du déplacement des formants.

2- Le modèle de production de parole.

Le modèle de production des sons voisés choisi pour référence est décrit par la figure 1.

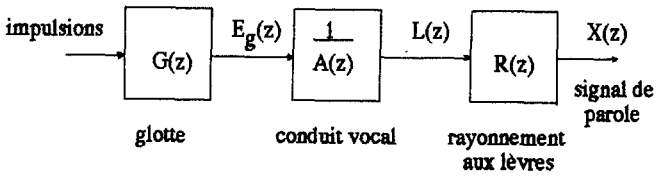

Fig. 1 Modèle de production des sons voisés.

La quasi-stationnarité du signal est admise. L'onde de débit $E_g(z)$ est le résultat du filtrage par G(z) d'impulsions espacées de la durée du pitch. Le fonctionnement du conduit vocal est modélisé par un filtre tout-pôle 1/A(z), R(z) est classiquement une différentiation qui exprime la relation entre l'onde de débit aux lèvres L(z) et l'onde de pression en champ libre X(z). Par convention les grandeurs caractéristiques des fonctionnements en présence d'air à la pression atmosphérique ou de mélanges hyperbares seront indicées respectivement par "a" ou "h". A toutes fins de simplifications -et sans induire de fortes contradictions avec la théorie [1]- les égalités $G_h(z) = G_a(z)$ et $R_h(z) = R_a(z)$ seront admises.

Cette modélisation suggère un traitement du signal en 3 phases:

- caractérisation d'un filtre et d'une excitation qui modélisent l'appareil phonatoire "hyperbare".
- déduction d'un filtre modifié qui modélise le conduit vocal "air atmosphérique".
- synthèse du signal corrigé par filtrage de l'excitation par le filtre modifié.

La théorie de la prédiction linéaire autorise cette approche : 2 formulations en ont été données. L'une par Makhoul dans l'hypothèse d'une correction spectrale non linéaire [5], l'autre par Schaffer dans l'hypothèse d'une

correction spectrale linéaire [6]. La formulation que nous proposons est une synthèse des 2 précédentes qui exploite avantageusement la décomposition de la déformation de la fonction de transfert du conduit vocal en 2 composantes: l'une linéaire, l'autre non linéaire.

3- Formulation de la déformation.

Il est généralement admis que la loi de Fant -Lindquist [7] justifiable pour le premier formant, qualifie de façon réaliste la transposition spectrale qui résulte de l'utilisation des mélanges de type héliox :

$$F_h^2 = (\frac{c_h}{c_a} F_a)^2 + (\frac{c_h}{c_a} F_{wa})^2 \, (\frac{\rho_h}{\rho_a} - 1) \qquad (1)$$

* notations:

F_a, F_h : fréquences

F_{wa} : fréquence de résonance du conduit vocal fermé aux lèvres.

c : vitesse du son.

ρ : masse volumique du gaz.

Une interprétation de (1) consiste à écrire que la fonction de transfert "air" se déduit point par point de la fonction de transfert "hyperbare" par 2 transformations successives:

1. calcul de

$$y(F_h) = \frac{c_h}{c_a} F_a$$

qui traduit la correction de la non linéarité introduite par le facteur

$(\frac{\rho_h}{\rho_a} - 1)$ dans la loi $F_h = f(F_a)$, soit:

$$y(F_h) = [\, F_h^2 - (\frac{c_h}{c_a} F_{wa})^2 \, (\frac{\rho_h}{\rho_a} - 1) \,]^{1/2} \qquad (2)$$

2. calcul de

$$F_a = \frac{c_a}{c_h} \, y(F_h) \qquad (3)$$

qui traduit la correction de la composante linéaire introduite par le rapport des vitesses du son c_h/c_a.

Compte-tenu de (2) et (3) la bande passante BP_a d'un formant se déduit de celle BP_h du formant transposé (de fréquence $F_h + BP_h/2$) par la loi :

$$BP_a = \frac{c_a}{c_h} [\, y(F_h + BP_h) - y(F_h) \,]$$

Pour les formants de fréquence élevée $y(F_h)$ est assimilable à F_h ; il vient alors:

$$BP_a \# \frac{c_a}{c_h} BP_h$$

et pour le premier formant:

$$BP_a < \frac{c_a}{c_h} BP_h$$

Le réalisme de ces résultats est controversé par certaines publications [8] : toutefois, dans le cadre de l'algorithme, une correction spécifique des bandes passantes serait délicate à mettre en oeuvre.

L'efficacité de l'algorithme est aussi conditionnée par la validité de l'extension de la correction aux sons non voisés: une étude des fricatives [1] confirme, dans ce cas particulier, la pertinence de la correction.

4- L'algorithme de correction.

La figure 2 représente l'organigramme de la correction spectrale conforme aux lois (2) et (3).

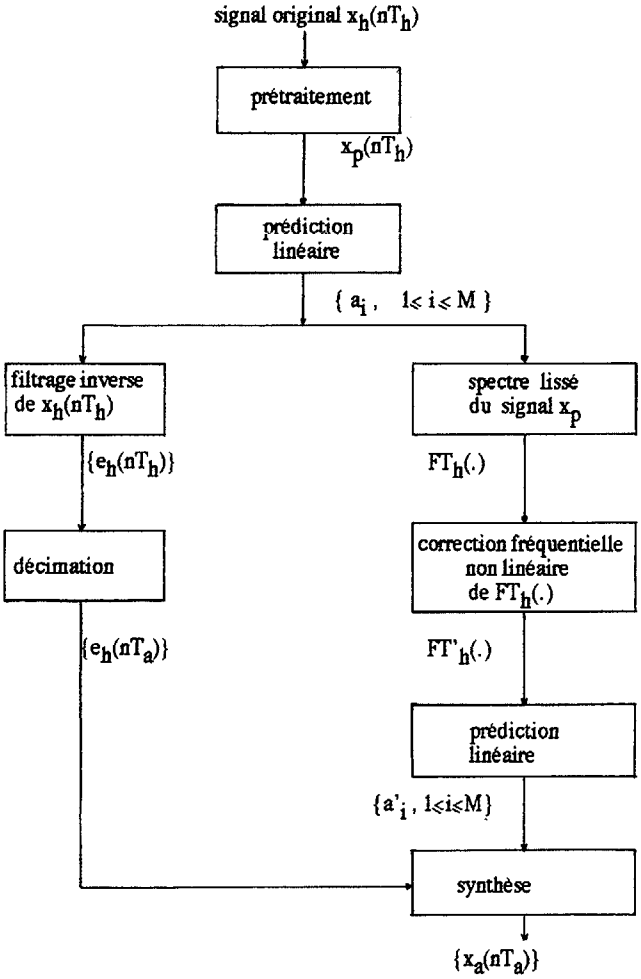

Fig. 2 Organigramme de la méthode de correction spectrale.

4-1 Analyse.

Le problème de la discrimination source-conduit vocal n'a de solution que si le modèle de production de parole est complété par une ou plusieurs hypothèses. Cette (ou ces) hypothèses ne peut être relative qu'à la source glottique dans la mesure où celle-ci est considérée indépendante de la configuration articulatoire : nous admettons $G_h(z)$ connu.

En conséquence l'analyse est effectuée comme suit :

- le signal de parole "hyperbare" $x_h(nT_h)$ échantillonné à la fréquence $1/T_h$ est pondéré par une fenêtre de Hamming et préaccentué par un filtrage de type $1-\mu z^{-1}$ dans le but de compenser l'effet de $G_h(z).R_h(z)$.

- le prédicteur d'ordre M $\{a_0, a_1, \ldots, a_M\}$ caractéristique de $A_h(z)$ est déterminé par la méthode d'autocorrélation [9]. Les variables d'entrée du calcul sont les M

premiers termes de la fonction d'autocorrélation du signal prétraité préalablement calculés par une méthode directe.

- le signal d'excitation $e_h(nT_h)$ est obtenu par filtrage inverse de $x_h(nT_h)$:

$$e_h(nT_h) = \sum_{i=0}^{M} a_i \, x_h[(n-i)T_h]$$

4-2 Caractérisation du filtre A'(z).

L'expression de 1/A'(z) définit la fonction de transfert $FT'_h(.)$ du conduit vocal pour un mélange respiratoire hyperbare fictif tel que $\rho_h = \rho_a$. Le prédicteur noté $\{a'_0, a'_1, \ldots, a'_M\}$ est obtenu par la méthode suivante :

- N points de la fonction de transfert fictive $FT'_h(.)$ régulièrement espacés de $1/NT_h$ sont calculés par :

$$FT'_h(n) = \frac{1}{|A(e^{j2\pi y^{-1}(\frac{n}{NT_h})})|^2} \qquad (4)$$

sachant que $y^{-1}(n)$ définit

$$F_h = [(\frac{n}{NT_h})^2 + (\frac{c_a}{c_h} F_{wa})^2 (\frac{\rho_h}{\rho_a} - 1)]^{1/2}$$

- la tranformée de Fourier inverse de $FT'_h(.)$ donne la fonction d'autocorrélation de la réponse impulsionnelle du conduit vocal "hyperbare" affranchi de la distorsion non linéaire : les M coefficients a'_i en résultent par application de la méthode d'autocorrélation.

4-3 Synthèse du signal corrigé.

La fonction de transfert du conduit vocal "air atmosphérique" se déduit de $FT'_h(.)$ par une homothétie des fréquences dans le rapport c_a/c_h : 1/A'(z) caractérise en conséquence le filtre de synthèse, moyennant une période d'échantillonnage T_a à la synthèse telle que :

$$T_a = T_h \frac{c_a}{c_h}$$

L'équation de synthèse s'écrit alors :

$$x_a(nT_a) = e_h(nT_a) - \sum_{i=1}^{M} a'_i \, x_a[(n-i)\,T_a]$$

sachant que le signal discret $e_h(nT_a)$ se déduit du signal continu $e_h(t)$ défini théoriquement par :

$$e_h(t) = \sum_{n=-\infty}^{+\infty} e_h(nT_h) \, \frac{\sin \frac{\pi}{T_h}(t - nT_h)}{\frac{\pi}{T_h}(t - nT_h)} \qquad (5)$$

4-4 Discussion de l'algorithme.

La principale originalité de l'algorithme tient au traitement de la composante linéaire de la déformation spectrale du signal de parole par un changement de fréquence d'échantillonnage: trois justifications en montrent l'intérêt :

- si un signal de parole "hyperbare" est défini dans un intervalle de fréquence]0, x KHz], son homologue "air" est pratiquement défini dans l'intervalle]0, x c_a/c_h KHz]: le changement de fréquence d'échantillonnage assure l'homogénéité de l'analyse et de la synthèse .

- à la synthèse, le nombre d'échantillons calculés est réduit dans le rapport c_h/c_a : argument important lorsque le temps réel est un objectif.

- la structure de l'algorithme rend par principe optionnel le traitement de la composante non linéaire : ce qui concrètement donne la possibilité de construire des séries d'appareils correcteurs moins onéreux mais néanmoins adaptés aux plongées de moyennes profondeurs.

5- Implémentation temps-réel de l'algorithme.

5-1 Optimisations logicielles.

Les spécifications matérielles et logicielles d'un processeur de signal imposent une optimisation de l'algorithme théorique. En l'occurence l'optimisation porte sur 3 points .

. La détermination d'un prédicteur.

Les coefficients de prédiction a_i sont solution d'un système de M équations à M inconnues de la forme

$$\sum_{i=1}^{M} a_i \, r|i-j| = -r(j)$$

où r(.) désigne la fonction d'autocorrélation discrète de la réponse impulsionnelle du conduit vocal [9]. Les particularités de ce système justifient divers algorithmes de résolution moins exigeants en volume de calcul que la solution générale , mais à priori non adaptés au calcul en virgule fixe. En conséquence l'algorithme de Leroux-Gueguen est retenu [10], bien que la solution caractérise le filtre par un jeu de coefficients de réflexion $\{K_1, K_2, \ldots, K_M\}$: une procédure simple assure le passage aux coefficients a_i [9].

. La détermination du filtre A'(z).

La connaissance du prédicteur $\{a_0, a_1, \ldots, a_M\}$ est l'hypothèse de la détermination. Une solution de calcul rapide de $FT'_h(.)$ doit être substituée à l'application de la relation (4). La fonction de transfert $FT_h(.)$ qui définit le fonctionnement du conduit vocal "hyperbare" est obtenue à partir d'une FFT de la séquence de N termes $\{a_0, a_1, \ldots, a_M, 0, \ldots, 0\}$. $FT'_h(n)$ pour $0 \leqslant n < N$ se calcule alors par une interpolation entre $FT_h(m)$ et $FT_h(m+1)$ pour

$$m = PE(y^{-1}(n)\,NT_h)$$

La transformée de fourier inverse de $FT'_h(.)$ produit la fonction d'autocorrélation discrète r'(.) de la réponse impulsionnelle du conduit vocal affranchi de l'effet de ρ_h : le prédicteur $\{a'_0, a'_1, \ldots, a'_M\}$ est solution du système :

$$\sum_{i=1}^{M} a'_i \, r'|i-j| = -r'(j)$$

. La détermination de l'excitation $e_h(nT_a)$.

L'application de la relation (5) n'est pas envisageable dans le traitement en temps réel: en conséquence une reconstitution du signal $e_h(t)$ par approximation polyno-

miale d'ordre 1 est effectuée :

$$e_h(t) = e_h(nT_h) + [e_h((n+1)T_h) - e_h(nT_h)] \frac{t - nT_h}{T_h}$$

$$\text{pour} \quad nT_h \leqslant t \leqslant (n+1)T_h$$

5-2. Architecture du système.

L'architecture du système multiprocesseurs (fig.3) qui a été conçu est de type MIMD (Multiple instruction streams/multiple data streams). Chaque processeur banalisé accède aux ressources communes via des bus spécialisés. Le signal d'entrée $x_h(t)$, échantillonné à une fréquence $c_h/(c_a T_a)$, est sectionné en fenêtres de Hamming de durée N T_h avec chevauchement de 20% . Les valeurs des paramètres N_h=500 et T_a=10 KHz, associées à un codage en 12 bits du signal, assurent un contexte technique favorable à l'algorithme.

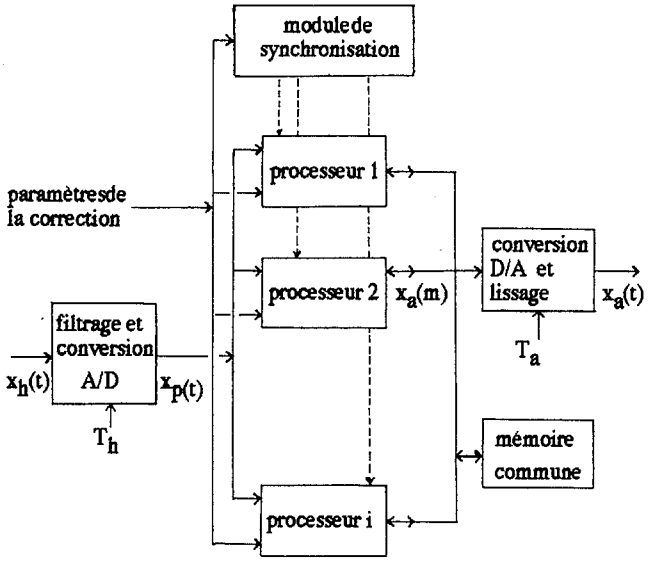

Fig. 3 Architecture du décodeur prototype.

Une implémentation temps-réel de l'algorithme a été évaluée en utilisant un processeur de première génération (TMS 32010). Les performances de ce processeur n'autorisent pratiquement que le traitement de la composante linéaire de la déformation spectrale. Trois processeurs peuvent assurer :

- le prétraitement du signal.
- le calcul des M premiers termes de la fonction d'autocorrélation (M=10).
- le calcul du prédicteur $\{a_0, a_1, \ldots, a_M\}$.
- la détermination de l'excitation $e_h(.T_a)$.
- la synthèse de la fenêtre de signal corrigé.

L'utilisation de processeurs de seconde génération est envisagée pour l'exécution en temps réel de l'algorithme dans son intégralité.

6- Conclusions.

Un algorithme et l'architecture d'un système approprié à son exécution en temps réel ont été étudiés. Une évaluation subjective de l'efficacité d'un appareil prototype a été établie par une série de tests d'intelligibilité à partir des listes de Griffiths [11]. Le résultat de ces tests est exclusivement significatif de la qualité des consonnes : un test consiste à reconnaitre un mot de type c-v-c parmi 4 mots phonétiquement proches. En l'occurence le taux moyen de reconnaissance de mots-tests, enregistrés à diverses profondeurs comprises entre 100 et 300 m, passe de 40% pour les mots déformés à 67% pour les mots corrigés. Ces chiffres ont seulement une valeur indicative.
Les tests de l'appareil ont été complétés au cours de la plongée expérimentale Hydra VI (Comex Déc. 86). La qualité de la parole décodée (intelligiblité, agrément, fidélité) s'est avérée bonne aux profondeurs inférieures à 450 m, mais s'est sensiblement détériorée au-delà : on peut présumer que dans ces conditions la correction de la composante non linéaire de la déformation spectrale du signal est nécessaire.

Références

[1] J. Crestel, M. Guitton : Thèses d'université à paraitre (Université de Rennes)
Laboratoire d'analyse des Systèmes de Traitement de l'Information.

[2] E. O. Belcher : Helium speech enhancement by frequency-domain processing. ICASSP 83, Boston.

[3] G. Duncan, M. A. Jack : Residualy excited LPC processor for enhancing helium speech intelligibility. Electronics Letters. Vol. 19 n° 18, 1983.

[4] J. Zurcher : Le transcodeur "CNET" de la voix en atmosphère d'hélium. Notice technique TMA/ETA/24 (1974)

[5] J. Makhoul : Methods for nonlinear spectral distorsion of speech signals. Bolts Beranck and Newman Inc Cambridge, Mass. 02138

[6] M. A. Richards : Helium speech enhancement design using the short-time Fourier transform. (Thesis) Georgia Institute of Technology (1982)

[7] G. Fant, J. Lindquist : Pressure and gas mixture effects on diver's speech. STL-QPSR-1 1968 (Royal Inst. Tech. Sweden (1968) pp. 7-21

[8] E. O. Belcher, S. Hatlestad : Formant frequencies bandwiths and Q's in helium speech. J. Acoust. Soc. Am. Vol. 74, n°2, August 1983

[9] J. D. Markel, A. H. Gray, Jr
Linear prediction of speech. Springer Berlin Heidelberg New York 1976.

[10] J. Leroux, C. Gueguen : A fixed point computation of partial correlation coefficients. IEEE Trans. on Acous., Speech, and Signal Processing. June 1977

[11] H.B.Rothman, R.Geffard, H. Hollien, C.J. Lanbertsen Speech intelligibility at high helium-oxygen pressures. Undersea Biomed. Res., Vol.7, n°4, pp.265-275, Dec.1980