TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

DEBRUITAGE DE LA PAROLE POUR LES RADIO-MOBILES

G. FAUCON, R. LE BOUQUIN

LABORATOIRE TRAITEMENT DU SIGNAL ET DE L'IMAGE / IRISA - UNIVERSITE DE RENNES I
CAMPUS DE BEAULIEU - 35042 RENNES CEDEX

RÉSUMÉ

Nous traitons le problème du débruitage de la parole pour les radio-mobiles en vue de la transmission du signal vocal. Nous avons proposé un certain nombre de méthodes. Nous effectuons ici une synthèse des méthodes qui se sont avérées satisfaisantes (bon compromis distorsion - réduction de bruit). Dans la situation à un seul microphone, deux modifications importantes sont apportées à la soustraction spectrale généralisée ; un filtre de Wiener avec contraintes spectrales est également présenté. Dans le cas de deux microphones, trois méthodes sont proposées, dont deux sont propres au cas de bruits spatialement décorrélés. Une comparaison de ces méthodes est donnée.

ABSTRACT

We deal with the enhancement of noisy speech for mobile radio applications. We already proposed some methods which have been performed on real noisy speech signals. We present a synthesis of the more efficient methods (compromise between distortion and noise cancellation). In the single microphone situation, we brought two modifications on the generalized spectral subtraction. An iterative Wiener filtering with spectral constraints is also presented. If two microphones are available, we propose three methods, two of which are developed for decorrelated noises. Finally, a comparison of these methods is given.

I. INTRODUCTION

Nous traitons ici le problème de la réduction du bruit sur la parole pour les radio-mobiles. Le combiné téléphonique utilisé dans une voiture sera à plus ou moins long terme remplacé par un téléphone mains-libres, et ce notamment pour des raisons de sécurité. Aussi, le rapport signal à bruit se trouve diminué et une opération de débruitage est nécessaire dans le cas d'une transmission du signal vocal pour accroître l'intelligibilité et la qualité du signal transmis, voire diminuer la fatigue de l'auditeur. Cette opération de débruitage peut également être utilisée en reconnaissance ; les méthodes alors retenues peuvent être différentes, les critères n'étant pas identiques pour la transmission et la reconnaissance.

Le but de ce papier est de présenter une synthèse des méthodes que nous avons nous-mêmes proposées et qui apportent réellement un gain afin de dresser un premier bilan [1,2]. D'autres méthodes que nous avons proposées se sont avérées insatisfaisantes en raison d'une distorsion trop importante ou d'une réduction de bruit jugée insuffisante. En particulier, les méthodes à deux voies telles que l'algorithme de Frost (à contrainte dure) et l'algorithme de Kaneda (à contrainte douce) incluant un prétraitement permettant de rendre les signaux identiques, les structures ISAB et IBRD [1] (où un premier étage permet, par identification de la fonction de transfert entre signaux seuls ou entre bruits seuls, de se ramener à un problème d'estimation d'un signal avec référence bruit seul ou signal seul) n'apportent pas une réduction de bruit suffisante.

Les tests subjectifs constituent le moyen d'évaluation le plus sûr. Même si nous n'avons pas conduit une campagne de tests d'écoute de façon intensive et rigoureuse, il est cependant possible de fournir des renseignements précieux pour chaque méthode. Des critères objectifs peuvent également être utilisés, tels que le gain sur le rapport signal à bruit ou diverses distances spectrales, lorsque signal et bruit sont enregistrés séparément. Le signal estimé est alors comparé au signal pur, l'erreur d'estimation incluant le bruit résiduel et la distorsion du signal. Cette distorsion linéaire introduite sur le signal et prise en compte lors des tests objectifs est plus ou moins préjudiciable à l'écoute. Inversement, des bruits particuliers induits par la méthode testée peuvent être désagréables à l'écoute sans être mis en évidence dans des tests objectifs.

Dans le paragraphe II, nous rappelons l'algorithme de la soustraction spectrale généralisée auquel nous apportons deux modifications puis nous introduisons des contraintes spectrales sur le filtre de Wiener. Nous examinons ensuite, au paragraphe III, la situation où deux microphones sont disponibles. Trois méthodes sont alors développées : deux sont basées sur une faible corrélation entre les bruits, la troisième ne fait aucune hypothèse a priori sur cette cohérence. Finalement, le paragraphe IV justifie les méthodes à retenir en raison de leur faible complexité en rappelant leurs hypothèses.

II. SITUATION A UN MICROPHONE

Dans le cas d'un seul microphone, nous recevons une observation x(t) constituée d'un signal, s(t), et d'un bruit, b(t), additifs et indépendants. Nous supposons disposer de périodes de "silence" (bruit seul) pour apprendre certaines caractéristiques du bruit.

a - soustraction spectrale modifiée

La soustraction spectrale est basée sur l'additivité et l'indépendance du signal et du bruit, et sur le fait que l'ouïe est insensible à la phase du signal. Ainsi, à partir du module du spectre du signal estimé et de la phase du signal bruité, nous pouvons reconstituer le signal par transformée de Fourier inverse :

$$s(t) = F^{-1}\left\{ |\hat{S}(f)|\, e^{j\theta_x} \right\} \qquad (1)$$

$|\hat{S}(f)|$ est obtenu en supposant qu'à court-terme, on puisse écrire :

$$|\hat{S}(f)|^2 = |X(f)|^2 - P_B(f) \qquad (2)$$

où X(f) est le spectre de l'observation x(t) sur le bloc considéré et $P_B(f)$ est la densité spectrale du bruit obtenue par moyennage sur les périodes de silence. $|\hat{S}(f)|^2$ est mis à 0 si la différence calculée en (2) est négative. Pour réduire le bruit musical apporté par cette méthode, Bérouti et al. [3] ont généralisé cet algorithme par l'introduction de trois paramètres. Nous calculons d'abord :

$$D(f) = |X(f)|^{2\delta} - \alpha.P_B^{\delta}(f)$$

puis nous déduisons :

$$|\hat{S}(f)| = \begin{cases} D(f)^{1/2\delta} & \text{si } D(f) \geq (\beta P_B(f))^{\delta} \\ (\beta P_B(f))^{1/2} & \text{sinon} \end{cases} \quad (3)$$

$\hat{s}(t)$ est ensuite obtenu par FFT inverse selon (1). Une autre façon de retrouver $\hat{s}(t)$ est d'appliquer un filtre à l'observation x, filtre de gain $H(f) = \hat{S}(f) / X(f)$ où $\hat{S}(f)$ est obtenu par l'opération de soustraction spectrale décrite en (3) et X(f) représente une estimée du spectre de l'observation. Donnons quelques précisions sur les paramètres : $\delta = 1$ correspond à une soustraction des d.s.p. et $\delta = 0.5$ à une soustraction d'amplitude. α est un facteur de surestimation du bruit et β permet, conjointement avec la d.s.p. du bruit apprise, de régler le plancher spectral. Les valeurs de δ et α que nous avons retenues et qui ont été reconnues optimales par d'autres chercheurs sont : $\delta = 0.5$, $\alpha = 1.5$. D'autre part, nous avons choisi $\beta = 10^{-4}$. A cette dernière méthode que nous retenons pour la suite, méthode associant le filtrage et l'algorithme de soustraction spectrale, nous apportons deux modifications :

i) itérations du facteur de surestimation

Théoriquement, le paramètre α devrait être égal à 1 et la quantité D(f) devrait toujours rester positive. Or, α est pris supérieur à 1 afin de réduire le bruit large bande et le bruit musical. La méthode est basée sur le fait que D(f) peut expérimentalement devenir négatif. Il apparaît alors intéressant de décrémenter α afin que D(f) reste non seulement positif mais aussi supérieur au plancher spectral $(\beta P_B(f))^{\delta}$.

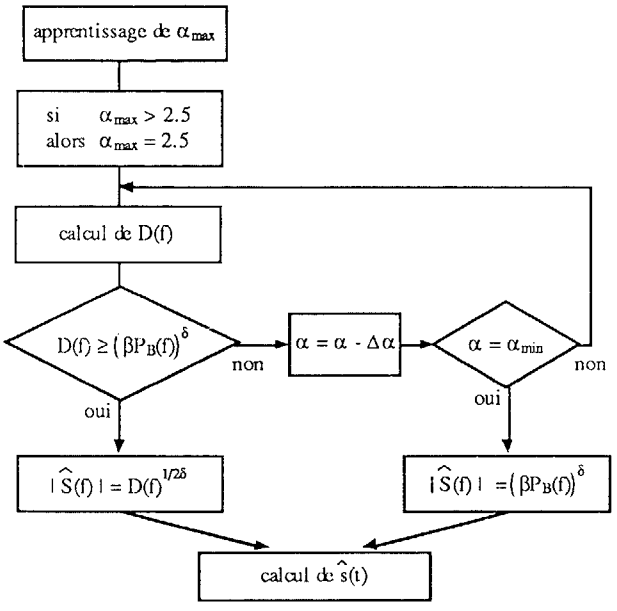

Figure 1 : Bloc-diagramme de la soustraction spectrale modifiée

Cette décrémentation de α sera faite en fonction des résultats de la comparaison de D(f) au plancher spectral [2]. α est initialisé par α_{max} évalué en absence de parole [4,5]. Cette grandeur est majorée par 2.5 et la valeur minimale de α est fixée à 0.5. L'algorithme correspondant est donné Figure 1.

ii) segmentation

Les traitements fréquentiels décrits travaillent par bloc de n échantillons et ne prennent pas en compte la stationnarité du signal sur plusieurs blocs consécutifs. Aussi, il apparaît intéressant de segmenter le signal afin de mieux connaître ses caractéristiques locales [6]. Des essais effectués par R. André-Obrecht ont montré que l'opération de segmentation pouvait être appliquée à la parole bruitée sans perdre trop d'information. La Figure 2 montre les résultats de la segmentation appliquée au signal vocal original et au même signal auquel on a ajouté du bruit de voiture. Nous remarquons que certaines frontières délimitant les segments sont perdues lorsque l'algorithme est appliqué à la parole bruitée (ceci s'explique par le fait que le bruit masque alors la parole). Les frontières restantes correspondent approximativement à celles obtenues sur le signal pur. Les résultats de la segmentation seront utilisés pour obtenir une meilleure estimée de $|X(f)|$. Nous découpons le signal par bloc de n échantillons se recouvrant à 50%. Nous considérons qu'un bloc appartient à un segment, si ce dernier contient au moins n/2 + 1 échantillons de ce bloc. Deux méthodes d'obtention du spectre pour le segment traité ont été proposées [2]. La première prend en compte tout le segment pour obtenir l'estimée de $|X(f)|$ et possède donc un délai parfois important à la transmission.

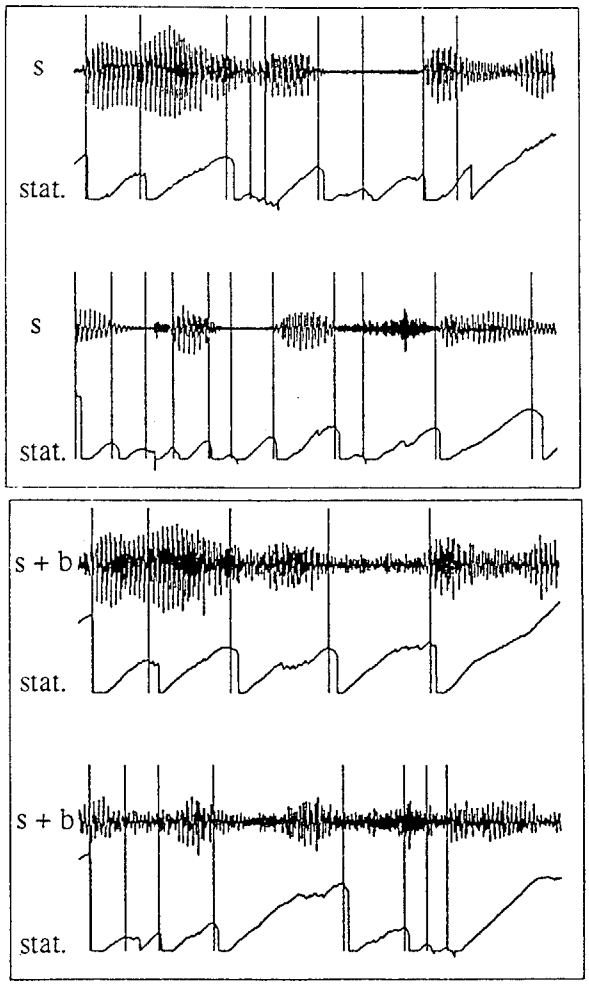

Figure 2 : Segmentation du signal pur et du signal bruité

Dans la deuxième solution proposée, nous nous imposons d'obtenir le signal avec un retard minimum. Pour chaque bloc p, nous calculons $|X_p(f)|$ par FFT. La valeur $\overline{|X_p(f)|}$ utilisée pour ce bloc sera obtenue de façon récursive :

$$\overline{|X_p(f)|} = (1 - \lambda^r).\overline{|X_{p-1}(f)|} + \lambda^r.|X_p(f)|$$

où r indique le numéro du bloc dans le segment considéré et λ est un facteur d'oubli inférieur à 1. Cette valeur $\overline{|X_p(f)|}$ sera utilisée aussi bien dans l'algorithme de soustraction spectrale pour déduire $|\hat{S}(f)|$ que dans l'opération de filtrage.

b - filtre de Wiener avec contraintes spectrales

Nous nous proposons maintenant d'estimer le signal d'après un filtrage de Wiener itératif bénéficiant d'une part de la mise à jour de la densité spectrale du signal au cours des itérations et d'autre part de l'application de contraintes spectrales.

La transformation LSP peut être vue comme une alternative de l'analyse LPC. Les coefficients LSP sont obtenus à partir d'une combinaison des polynômes prédicteurs direct et rétrograde (respectivement A(z) et B(z)) associés au signal de parole :

$$P(z) = A(z) + B(z) \quad \text{et} \quad Q(z) = A(z) - B(z)$$

Les angles des racines de P(z) et Q(z), $\{\omega_i, i = 1, 2, ..., M\}$ sont appelés les coefficients LSP. Ceux-ci possèdent d'importantes propriétés qui justifient leur utilisation dans l'application de contraintes spectrales. D'autre part, ces coefficients ont une signification importante au niveau de la perception.

Deux classes de contraintes, inter-trames et intra-trames, ont été appliquées sur le signal estimé [7]. Seules les contraintes inter-trames apportent une réelle amélioration. Celles-ci consistent en un lissage des coefficients de position (angles des racines de P(z)) au cours du temps en utilisant une fenêtre triangulaire de support variable. Les fréquences de formants les plus basses sont lissées sur une fenêtre de largeur plus faible que les fréquences plus élevées. Pour obtenir le coefficient lissé $\mathbb{P}_n(i)$ au bloc n, nous calculons les coefficients de position P(i) (i = 1, 2, 3, 4) pour les blocs n-3 à n+3. Pour i donné, si la majorité des valeurs de P(i) est située dans le premier quadrant, alors :

$$\mathbb{P}_n(i) = [\, 3P_n(i) + 2P_{n-1}(i) + 2P_{n+1}(i) + P_{n-2}(i) + P_{n+2}(i) \,] / 9$$

Sinon, le lissage se fait sur sept blocs de la manière suivante :

$$\mathbb{P}_n(i) = [\, 4P_n(i) + 3P_{n-1}(i) + 3P_{n+1}(i) + 2P_{n-2}(i) + 2P_{n+2}(i) + P_{n-3}(i) + P_{n+3}(i) \,] / 16$$

Aucun lissage n'est effectué sur les coefficients de différence (obtenus par différence des angles des racines de P(z) et de Q(z)). Cependant, il arrive qu'un coefficient de différence prenne une valeur trop faible, si bien que la valeur de $|d_i|$ est mise à d_{min}. Les angles des racines de Q(z) sont alors déduits des coefficients de position lissés et des coefficients de différence. A partir des polynômes P(z) et Q(z), on obtient le polynôme prédicteur A(z) = (P(z) + Q(z)) / 2. On obtient une densité spectrale du signal, ce qui permet d'effectuer un nouveau filtrage.

III. SITUATION A DEUX MICROPHONES

Nous examinons maintenant la situation où deux microphones sont disponibles. Les observations s'écrivent $x_1 = s_1 + b_1$ et $x_2 = s_2 + b_2$. Nous supposons que le signal à estimer est le signal s_1. La fonction de cohérence entre signaux de parole reste proche de 1 en module quelle que soit la fréquence. Pour les deux premières méthodes présentées, la distance entre microphones est choisie suffisamment grande pour supposer que les bruits sont décorrélés.

• La première méthode est basée sur une mesure de la cohérence entre observations [2]. Cette fonction constitue un critère pertinent pour savoir si, à une fréquence donnée, existe un signal de parole ou non. Les hypothèses de base sont : $|\rho_{s_1 s_2}| = 1$ et $|\rho_{b_1 b_2}| = 0$. Si la fonction de cohérence entre observations avoisine 1 en module, cela signifie que le bruit est négligeable vis-à-vis du signal ; si $|\rho_{x_1 x_2}|$ est proche de 0, cela signifie que le bruit est prépondérant. L'idée est de filtrer l'observation par une grandeur fonction du module de la cohérence. On définit la quantité MSC(f) (Magnitude Squared Coherence) comme $MSC(f) = |\rho_{x_1 x_2}(f)|^2$. L'algorithme utilisé est le suivant :

si $MSC(f) < S_{min}$ alors $\hat{S}_1(f) = S_{min}^{\alpha}.X_1(f)$

si $MSC(f) > S_{max}$ alors $\hat{S}_1(f) = X_1(f)$

si $S_{min} < MSC(f) < S_{max}$ alors $\hat{S}_1(f) = MSC(f)^{\alpha}.X_1(f)$

où α est un paramètre à choisir à la suite de tests d'écoute pour obtenir le meilleur compromis "distorsion-réduction de bruit". La fonction de cohérence entre observations est obtenue de la façon suivante : sur chaque bloc p, on calcule les transformées de Fourier des observations x_1 et x_2, notées $X_1(f,p)$ et $X_2(f,p)$ puis on déduit :

$$\rho_{x_i x_j}(f,p) = \beta.\rho_{x_i x_j}(f,p-1) + (1-\beta).X_i(f,p)X_j^*(f,p)$$

$$(i,j) = (1,2), (1,1), (2,2)$$

• La deuxième méthode est basée sur le filtrage de Wiener vectoriel, également appliqué dans le cas de bruits supposés décorrélés. Celui-ci nécessite la connaissance des densités spectrales des signaux et des bruits. Nous posons le problème du maximum de vraisemblance pour déterminer les paramètres nécessaires à ce filtrage. En réalité, on ne dispose que des observations x_1 et x_2. L'algorithme EM (Estimate-Maximize) est alors utilisé pour trouver les paramètres de manière itérative :

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta} \in \Theta} E\left\{ \log f_{\underline{Y}}(\underline{y} ; \underline{\theta}) / \underline{x} ; \underline{\theta}^{(n)} \right\} \qquad (4)$$

où $f_{\underline{Y}}(\underline{y} ; \underline{\theta})$ est la densité de probabilité du vecteur de données complètes $\underline{Y}$ et $E\left\{./ \underline{x} ; \underline{\theta}^{(n)} \right\}$ est la moyenne conditionnelle à $\underline{x}$ (vecteur des données incomplètes), calculée en utilisant le vecteur paramètre $\underline{\theta}^{(n)}$ [8]. L'idée qui mène à l'équation (4) est de choisir $\underline{\theta}$ qui maximise $\log f_{\underline{Y}}(\underline{y} ; \underline{\theta})$; puisque les données complètes ne sont pas disponibles, on maximise son espérance conditionnellement aux valeurs observées $\underline{x}$. Cette moyenne conditionnelle n'est pas exacte puisque $\underline{\theta}^{(n)}$ ne représente pas la valeur réelle de $\underline{\theta}$ mais une estimée courante. Ainsi, l'algorithme itère, utilisant chaque nouvelle estimée du vecteur paramètre pour améliorer la moyenne conditionnelle sur la prochaine itération et ainsi l'estimée du vecteur paramètre suivante. Les données complètes que nous avons choisies sont x_1, x_2, s. Le maximum de vraisemblance répond à l'équation :

$$L_c(\underline{\theta}) = \log f_{x_1,x_2,s}(x_1(t),x_2(t),s(t) ; \underline{\theta})$$

et comme les signaux $x_1(t)$ et $x_2(t)$ sont indépendants, à s(t) donné, on obtient :

$$L_c(\underline{\theta}) = \log f_s(s(t) ; \underline{\theta})$$

$$+ \log f_{x_1/s}(x_1(t)/s(t) ; \underline{\theta}) + \log f_{x_2/s}(x_2(t)/s(t) ; \underline{\theta})$$

Maximiser la vraisemblance des données complètes par rapport à $\underline{\theta}$ est équivalent à maximiser chaque terme par rapport aux paramètres dont il dépend. En [2,9], nous donnons des renseignements complémentaires sur la mise en œuvre de cet algorithme.

• La dernière méthode que nous décrivons ici est la méthode P.I.S. [1] (Prétraitement + Identification entre Signaux), qui ne fait a priori aucune hypothèse forte sur la corrélation entre bruits. Celle-ci est représentée Figure 3. Il semble cependant préférable que la corrélation entre bruits reste faible. Dans cette structure, nous effectuons une réduction de bruit sur chaque voie (par filtrage de Wiener, soustraction spectrale, ...). Nous obtenons ainsi deux nouvelles observations x'_1 et x'_2 constituées chacune du signal de parole filtré et d'un bruit résiduel que l'on écrit sous la forme : $x'_1 = s'_1 + b'_1$ et $x'_2 = s'_2 + b'_2$. Celles-ci présentent un meilleur rapport signal à bruit qu'en x_1 et x_2. Il est alors possible, à partir de la référence x'_2, d'obtenir une estimée du signal s_1, ou du signal s'_1, suivant la configuration choisie. x'_2 est envoyé sur un filtre dont la sortie est soustraite d'une voie, dite voie principale, choisie comme étant x_1 ou x'_1, le résultat étant utilisé pour adapter les coefficients du filtre afin de minimiser la puissance en sortie du soustracteur. La sortie du filtre, fournissant une estimée du signal est additionnée à la grandeur x'_1, et le résultat de cette somme est divisé par deux pour fournir l'estimée finale du signal. Une étude théorique de cette structure lorsque les filtres utilisés sont optimaux et lorsque les bruits sont décorrélés est donnée en [10].

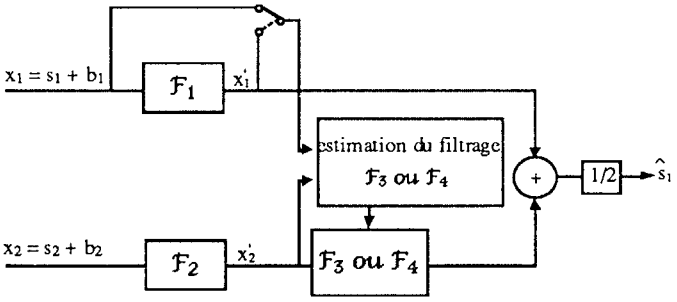

Figure 3 : Structure P.I.S.

IV. RESULTATS ET CONCLUSION

Nous avons testé ces différentes méthodes sur des fichiers réels de parole bruitée, enregistrée dans une voiture en roulement, à vitesse stabilisée (bruits quasiment stationnaires). Des résultats objectifs tels que le gain sur le rapport signal à bruit ou les distances spectrales ne peuvent être évalués que lorsque signal et bruit ont été enregistrés séparément. A titre indicatif, nous donnons une courbe de gain sur le rapport signal à bruit (Figure 4) pour la méthode basée sur la cohérence, lorsque les deux signaux sont identiques puis différents. Les méthodes présentées ici apportent une amélioration substantielle lors d'écoutes du signal estimé. Une cassette audio permet de présenter les résultats obtenus par les différents traitements. Les méthodes que nous retenons en raison de leur efficacité et de leur simplicité sont :

• la soustraction spectrale avec itérations du facteur de surestimation : elle nécessite l'apprentissage des caractéristiques du bruit et sa stationnarité à court-terme.

• la méthode basée sur la cohérence : elle nécessite au moins deux voies, et la corrélation entre bruits doit être négligeable.

• la méthode P.I.S. : elle nécessite l'apprentissage des caractéristiques du bruit ; elle prend en compte le changement de position du conducteur.

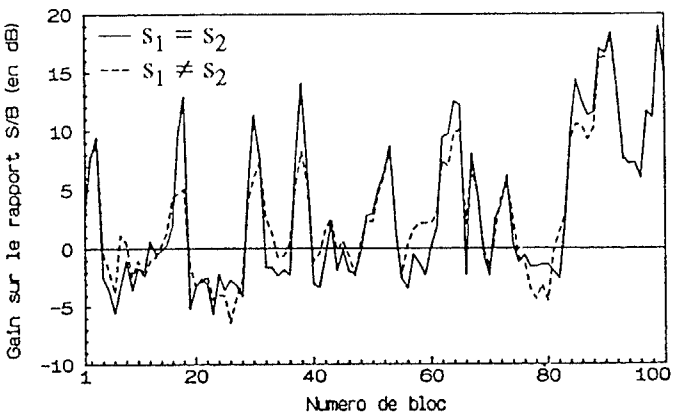

Figure 4 : Gain sur le rapport signal à bruit

Ces trois méthodes ont une complexité raisonnable, et leur implantation sur un processeur est possible. A titre d'exemple, le temps moyen pour traiter un échantillon est de 40 µs (à comparer à la période d'échantillonnage : 125 µs) sur un processeur AT&T DSP 16. Ces méthodes restent cependant à valider dans des environnements de bruits non stationnaires.

Les autres méthodes que nous avons proposées, si elles apportent une amélioration sur le signal bruité, sont de complexité plus importante : filtrage de Wiener vectoriel, débruitage après segmentation, filtrage de Wiener avec contraintes spectrales. D'autre part, ces deux dernières restituent le signal avec un retard qui peut être important.

REFERENCES

[1] S. TAZI MEZALEK, "Algorithmes de Débruitage de la Parole pour les Radio-Mobiles", Thèse de Doctorat de l'Université de Rennes I, Septembre 1990.

[2] R. LE BOUQUIN, "Traitements pour la Réduction du Bruit sur la Parole. Application aux Communications Radio-Mobiles", Thèse de Doctorat de l'Université de Rennes I, Juillet 1991.

[3] M. BEROUTI et al., "Enhancement of Speech Corrupted by Acoustic Noise", ICASSP, pp. 208-211, April 1979.

[4] G. FAUCON et al., "Bilan sur Trois Méthodes de Débruitage de la Parole", Séminaire Traitement et Représentation du Signal de Parole, Le Mans, Juin 1991.

[5] P. LOCKWOOD, J. BOUDY, "Experiments with a Non-Linear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars", à paraître dans EUROSPEECH 91.

[6] R. ANDRE-OBRECHT : "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals", IEEE on ASSP, vol. 36, n°1, pp. 29-40, January 1988.

[7] J.H.L. HANSEN, M.A. CLEMENTS, "Constrained Iterative Speech Enhancement with Application to Speech Recognition", IEEE, Trans. on Signal Processing, vol. 39, n°4, pp. 795-805, April 1991.

[8] M. FEDER, A.V. OPPENHEIM, E. WEINSTEIN, "Maximum Likelihood Noise Cancellation Using the EM Algorithm", IEEE on ASSP, vol. 37, n°2, pp. 204-216, February 1989.

[9] R. LE BOUQUIN, G. FAUCON, "Maximum Likelihood Noise Cancellation with Spectral Constraints", ICASSP 1991, Toronto, pp. 941-944, May 1991.

[10] G. FAUCON, S. TAZI MEZALEK, "Structures de Wiener et Structure P.I.S. pour l'Estimation d'un Signal", Annales de ce Colloque.

Les auteurs remercient la société MATRA COMMUNICATION et le CNET de Lannion A pour les enregistrements fournis.