Réflexion sur un annuleur d'écho robuste aux non stationnarités de la parole

S. Larbi[1] M. Jaidane[1] M. Turki[1] M. Bonnet[2]

[1]Laboratoire des Systèmes de Communications – ENIT BP 37, 1002 Tunis, Tunisie

[2] UFR Math-Info, Université René Descartes–Paris V – France

slarbi@excite.com – monia.turki@enit.rnu.tn

Madeleine.Bonnet@math-info.univ-paris5.fr – meriem.jaidane@enit.rnu.tn

Résumé

Dans les systèmes de prise de son mains-libres, les annuleurs d'écho adaptatifs souffrent en général, d'un manque de robustesse aux non stationnarités du signal de parole. Selon le schéma d'AEC[1] proposé, c'est un signal de parole "stationnarisé" qui pilote l'annuleur d'écho. Ce dernier est la résultante du signal de parole auquel se rajoute un signal inaudible et stationnaire par morceaux, provenant de la mise en forme spectrale d'un bruit blanc par un filtre perceptif issu du signal de parole. Une telle approche s'apparente à celle proposée dans [2] pour le contexte stéréophonique. Cette mise en forme spectrale est inspirée des techniques de tatouage audionumérique [1].

Mots-clé

Systèmes de prise de son mains-libres, annulation d'écho adaptative, stationnarisation du signal de parole, mise en forme perceptuelle.

1 Introduction

Dans les systèmes de prise de son mains-libres (audio- et visioconférence,...) la qualité de la transmission est altérée par le phénomène d'écho, découlant du couplage acoustique entre haut-parleurs et microphones. L'écho est classiquement réduit par un système dit d'annulation d'écho (AEC), destiné à identifier en temps réel les réponses impulsionnelles de couplage. Les algorithmes adaptatifs permettant cette identification présentent des problèmes de robustesse dus aux variations temporelles des caractéristiques statistiques du signal de parole (corrélation, moments d'ordre supérieur à 2,...).

Différentes approches ont été présentées, essentiellement pour remédier aux problèmes de variation de la puissance des signaux [3] (algorithmes LMS normalisés) et aux problèmes de corrélation de ces signaux qui limitent les performances transitoires et les capacités de poursuite des algorithmes (algorithmes de projection affine [3], structures préblanchissantes [4]).

Nous proposons ici, dans un contexte d'annulation d'écho monophonique, de piloter l'AEC par un signal dont les propriétés statistiques ont été modifiées afin d'améliorer les performances de l'AEC. Après avoir présenté cette nouvelle structure d'AEC, nous en étudierons les performances, en insistant sur les paramètres permettant de modifier les caractéristiques du signal de parole pilotant l'AEC.

[1]AEC : Acoustic Echo Canceller

2 Nouvelle structure d'AEC

2.1 Schéma proposé

Classiquement, dans un schéma d'AEC, le signal de parole x_n issu du haut-parleur et provenant du locuteur lointain est filtré par le chemin d'écho H. Le microphone reçoit l'écho y_n auquel s'ajoute la parole locale et le bruit ambiant b_n. C'est le signal x_n et l'erreur résiduelle e_n qui pilotent l'annuleur d'écho.

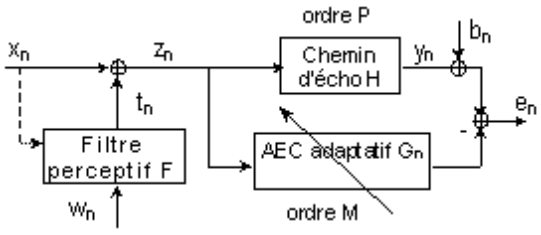

FIG. 1 – Schéma d'AEC proposé

Dans le schéma proposé figure 1, ce n'est plus le signal de parole x_n qui pilote l'AEC adaptatif G_n, mais le signal *modifié*,

$$z_n = x_n + t_n. \quad (1)$$

Les équations qui décrivent le fonctionnement de l'AEC de structure linéaire dans le cas de l'algorithme LMS normalisé sont :

$$\begin{cases} e_n = (y_n + b_n) - G_n^T Z_n \\ G_{n+1} = G_n + \mu_n e_n Z_n \end{cases} \quad (2)$$

où μ_n représente le pas d'adaptation normalisé par la puissance du signal z_n et $Z_n = (z_n, ..., z_{n-P)^T}$ est le vecteur observation.
Les performances de l'AEC sont mesurées dans ce qui suit par l'évolution de l'ERLE[2] Si l'on suppose que les filtres de l'AEC et du chemin d'écho ont même longueur, on peut également mesurer la qualité de l'annulation d'écho à travers la déviation :

$$V_n = H - G_n \quad (3)$$

On montre ainsi d'après (2) et (4) que

$$V_{n+1} = (I - \mu Z_n Z_n^T) V_n + \mu e_n b_n \quad (4)$$

L'ensemble des performances en régime transitoire et permanent est lié au caractéristiques statistiques de la matrice $R_n = (I - \mu Z_n Z_n^T)$. Ainsi, pour les faibles pas d'adaptation et dans un contexte stationnaire, on montre que plus la dispersion des valeurs propres de $E[Z_n Z_n^T]$ est faible, meilleure est la convergence. Ceci est en particulier obtenu lorsque le signal z_n est blanc. Dans le contexte réel où l'entrée x_n est non stationnaire, l'algorithme souffre d'un manque de robustesse vis à vis des statistiques de l'entrée. L'ajout du signal non audible t_n agit dans le sens de la stationnarisation du signal x_n.
C'est à partir du filtre perceptif F (figure 1) que le signal t_n est construit. La réponse fréquentielle de ce filtre est obtenue à partir du signal de parole x_n. Un signal stationnaire (bruit pseudo-blanc) w_n est mis en forme spectralement par un filtre F, dont les coefficients sont renouvelés toutes les 20 ms, pour générer un signal t_n non audible en présence de x_n, appelé signal de tatouage. C'est donc le nouveau signal z_n issu du haut-parleur qui génère l'écho y_n.

2.2 Synthèse du filtre perceptif F

Dans le cas général des signaux audio, la synthèse des filtres de mise en forme spectrale nécessite la mise en oeuvre de modèles psychoacoustiques. Nous considérons ici le signal de parole comme étant stationnaire sur des durées d'environ 20 ms et autorégressif à l'ordre Q (il est issu du filtrage d'un bruit blanc par un filtre de fonction de transfert tout-pôles $1/A(z)$).
Sachant que la courbe de masquage d'un signal audio épouse grossièrement sa DSP[3], on se contentera ici, comme cela est fait habituellement dans les codeurs

[2]ERLE : Error Return Loss Enhancement
[3]DSP : Densité Spectrale de Puissance.

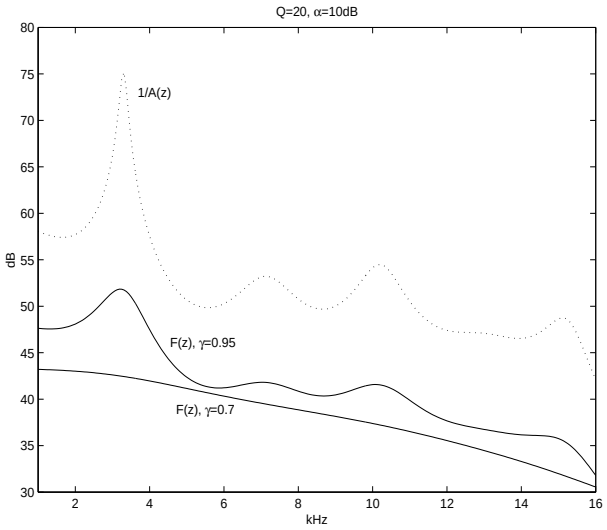

FIG. 2 – Réponses fréquentielles $F(z)$ du filtre perceptif (trait plein) et $1/A(z)$ (pointillés)

de parole [5], d'introduire un facteur de pondération $0 < \gamma < 1$ tel que

$$F(z) = 1/A(\frac{z}{\gamma}) \quad (5)$$

Cette pondération sert à mettre en forme le spectre du bruit w_n (cf. figure 1) de manière à ce qu'il soit masqué par le signal de parole x_n [6]. La figure 2 représente la réponse en fréquence du filtre $1/A(z)$ correspondant à 20 ms de parole échantillonnée à 32 kHz et la fonction de transfert du filtre F pour $\gamma = 0.95$ et 0.7. La réponse fréquentielle F(z) est généralement atténuée d'un facteur α, égal à 10 dB sur la figure 2, pour s'assurer de l'inaudibilité.

3 Performances de l'AEC proposé

3.1 Amélioration des performances

L'amélioration attendue des performances de l'AEC, du fait de la stationnarisation du signal de parole, concerne à la fois sa vitesse de convergence et sa capacité de poursuite. Quelques résultats de simulation sont donnés sur les figures 3 et 4. La figure 3 montre l'évolution en régime transitoire de la déviation quadratique $DQ = (H - G_n)^T (H - G_n)$ instantanée dans le cas classique et selon la nouvelle structure, pour un algorithme LMS normalisé et un même ordre des filtres H et G, ici M=P=300 (réponse impulsionnelle tronquée d'un chemin d'écho réel). La convergence la plus rapide a été effectivement obtenue avec le signal de parole tatoué z_n. Elle est plus rapide dans les zones non voisées de x_n. En effet, dans ces zones, le signal a un aspect bruité [7], il est donc moins corrélé que dans les zones voisées où il a un aspect harmonique.

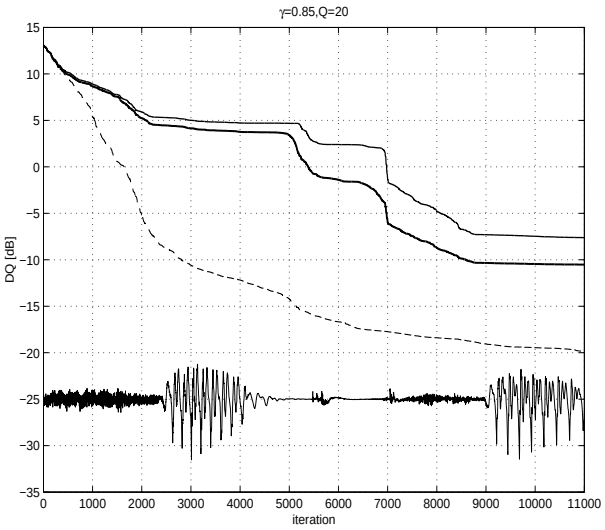

FIG. 3 – Déviation quadratique en régime transitoire pour l'AEC classique et proposé (en gras) et piloté par t_n uniquement (pointillés)

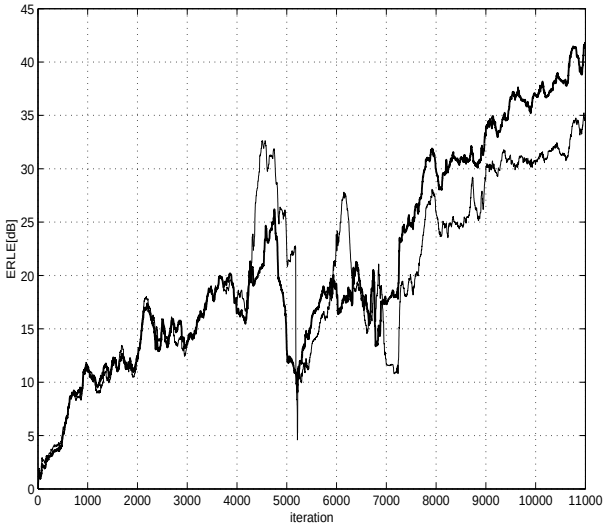

FIG. 4 – ERLE en régime transitoire pour l'AEC classique et proposé (en gras)

Afin de mieux comprendre la modification apportée par l'introduction d'un signal de tatouage, nous avons visualisé également sur la figure 3, l'évolution de la DQ lorsque l'AEC est piloté par le signal de tatouage t_n (courbe en pointillés). L'accélération de la vitesse de convergence obtenue avec l'AEC piloté par le seul signal t_n explique, à priori, l'amélioration dans le cas de la nouvelle structure. Sur la figure 4, l'amélioration apportée est également visible au niveau de l'ERLE, dont la valeur est calculée sur des fenêtres glissantes de taille $N = 256$ par :

$$ERLE(n) = \frac{\sum_{i=n-N+1}^{n} y_i^2}{\sum_{i=n-N+1}^{n} e_i^2} \qquad (6)$$

3.2 Influence des paramètres de la nouvelle structure

Le schéma d'AEC de la figure 1 possède par ailleurs quatre degrés de liberté, ce sont :
- l'ordre Q du filtre perceptif $F(z)$,
- les valeurs de la pondération γ et de l'atténuation α
- la nature du bruit w_n (gaussien, uniforme,...).

Nous avons fait varier sur la figure 5 le facteur de pondération γ pour un ordre $Q = 20$ et $\alpha = 10$ dB. Une valeur élevée de γ favorise l'apparition des harmoniques dans $F(z)$ et donc l'aspect voisé et corrélé du signal t_n, ce qui ralentit la convergence. Par contre, une augmentation de la puissance de t_n est observée, ce qui améliore la convergence dans les zones non voisées. Cependant, un γ trop proche de 1 n'assure plus l'inaudibilité. Le choix du facteur de pondération est donc un compromis entre vitesse de convergence et fidélité au signal de parole original.
L'atténuation spectrale α sert principalement à assurer plus encore l'inaudibilité de t_n. Plus elle est importante, plus t_n est faible en puissance et moins la convergence de l'AEC est rapide.
L'ordre Q agit sur la convergence dans les zones voisées, car plus il est grand et plus le caractère voisé de x_n est conservé dans t_n. Par conséquent, la vitesse de convergence est meilleure pour Q faible, cependant la qualité du masquage perceptuel nécessite une valeur minimale de Q. Ceci est illustré par la DQ instantanée pour $Q = 10$ et 50 sur la figure 6. A titre de comparaison, nous avons tracé sur la même figure la DQ dans le cas classique (en gras).
Nous avons également constaté l'influence des statistiques de w_n sur les performances de l'AEC. La figure 7 montre ainsi pour un fort pas d'adaptation ($\mu = 0.001$) qu'une entrée gaussienne est préférable à une entrée uniforme. En effet, dans le cas où w_n est uniforme il n'y a pas d'amélioration visible par rapport au cas classique.
Dans toutes les simulations des figures 2 à 7, nous avons pris P=M=300, un pas $\mu = 0.001$, un SNR=60 dB entre x_n et le bruit ambiant b_n, et un signal w_n gaussien (sauf pour la courbe en pointillés de la figure 7, o il est uniforme).

4 Conclusion

La qualité de l'annulation d'écho dans les systèmes de prise de son mains-libres est en particulier liée à la robustesse des algorithmes d'adaptation. L'approche proposée consiste à piloter l'annuleur d'écho par un signal "stationnarisé" selon une technique qui s'apparente aux techniques de tatouage audionumérique

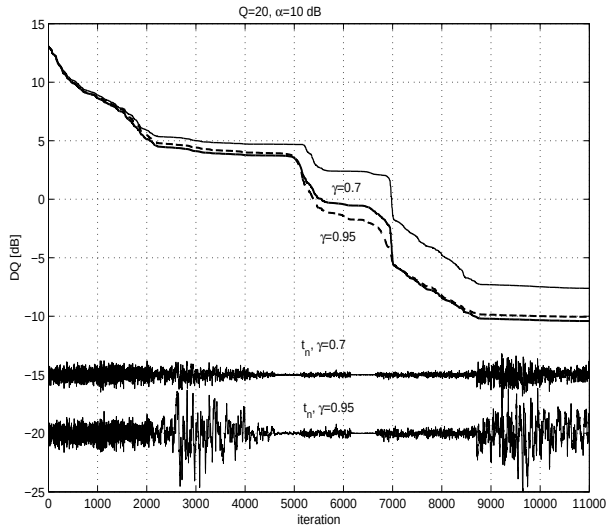

FIG. 5 – Influence du facteur de pondération sur la vitesse de convergence dans la structure proposée

[1]. Des résultats convaincants en terme de vitesse de convergence et de comportement en régime permanent ont été obtenus. Cette première réflexion doit être approfondie afin d'optimiser à la fois les paramètres du filtre perceptif et les caractéristiques du bruit mis en forme. Par ailleurs, cette étude doit être généralisée dans des contextes réels de double parole et de réponse impulsionnelle longue dans les systèmes d'annulation d'écho acoustique.

Références

[1] L. Boney, A. H. Tewfik, K. N. Hamdy "Digital watermarks for audio signals", IEEE Int. Conference on Multimedia Computing and Systems, Japan, June 1996.

[2] A. Gilloire, V. Turbin, "Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers", ICASSP 1998.

[3] S. Haykin, "Adaptive Filter Theory", Prentice-Hall, 1991.

[4] M. Mboup, "Identification adaptative par structures prédictives et récursives : application à l'annulation d'écho acoustique", PhD thesis, Paris-Sud, 1992.

[5] N. Moreau,"Techniques de compression des signaux", Masson, 1995.

[6] E. Zwicker, E. Feldtkeller, "Psychoacoustique, l'oreille récepteur d'information", Masson, 1981.

[7] R. Boite et M. Kunt, "Traitement de la parole", Presses Polytechniques Romandes, 1987.

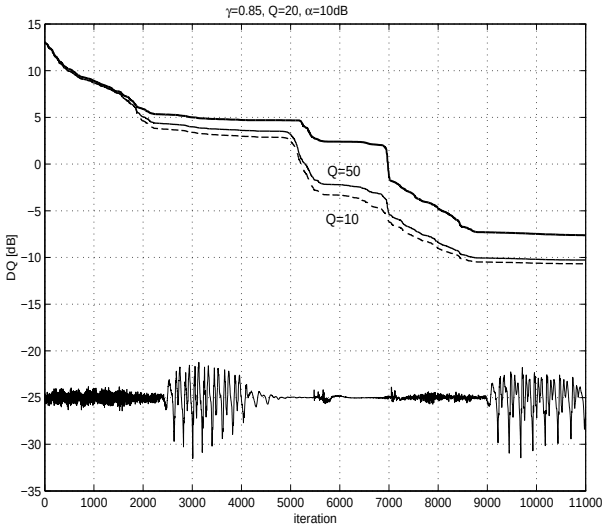

FIG. 6 – Influence de l'ordre Q de $F(z)$ sur la vitesse de convergence dans la structure proposée

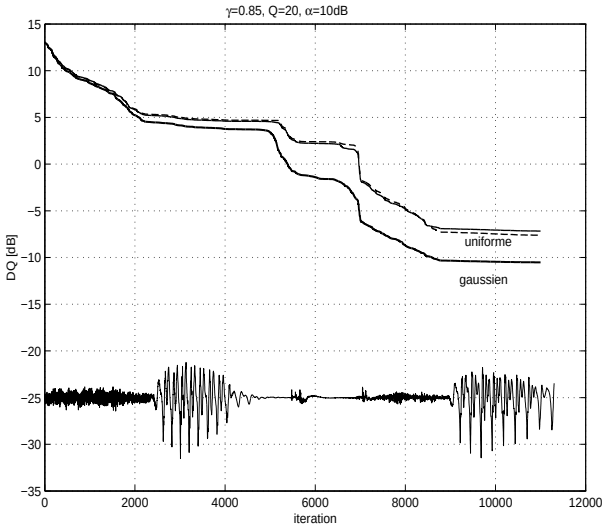

FIG. 7 – Influence de la nature du bruit w_n sur la vitesse de convergence dans la structure proposée