Amélioration psychoacoustique du filtrage de Wiener

Asmaa AMEHRAYE[1,2], Dominique PASTOR[1] et Ahmed TAMTAOUI[3]

[1]Ecole Nationale Supérieure des Télécommunications de Bretagne, CNRS (UMR 2872),
Technopôle de Brest-Iroise, CS 83818, 28238 Brest, France

[2]GSCM-LRIT, Université Mohammed V-Agdal
Faculté des Science de Rabat, Maroc

[3]Institut National des Postes et Télécommunications,
2, av Allal EL Fassi, Madinat AL Irfane, Rabat, Maroc

asmaa.amehraye,dominique.pastor@enst-bretagne.fr,
tamtaoui@inpt.ac.ma

Résumé – Dans ce papier, on s'intéresse à la réduction du bruit de type musical qu'engendrent des méthodes basées sur la soustraction de bruit et en particulier, le filtrage de Wiener. On compare plusieurs méthodes qui introduisent des modifications perceptuelles de ce filtrage et on propose une nouvelle méthode qui améliore la qualité de la parole débruitée en sortie du filtre de Wiener usuel. Cette amélioration résulte d'un contrôle du filtre de Wiener par un second filtre qui peut être considéré comme un facteur de pondération perceptuelle.

Abstract – This paper deals with musical noise resulting from subtractive type algorithms and especially Wiener filtering. We compare several methods that introduce perceptually motivated modifications of the standard Wiener filtering and we propound a new speech enhancement technique. This one aims to improve the quality of the enhanced speech signal provided by the standard Wiener filtering by controlling the latter via a second filter regarded as a psychoacoustically motivated weighting factor.

1 Introduction

Le débruitage de la parole en vue de l'amélioration de l'intelligibilité audio est un domaine de recherche très actif et présent dans de nombreux champs d'applications. Les méthodes classiques et largement utilisées se fondent sur la soustraction spectrale et le filtrage de Wiener. Ces méthodes parviennent à réduire efficacement le bruit additif. En contrepartie, elles produisent un bruit résiduel perceptuellement gênant et connu sous le nom de bruit musical. Les premières tentatives de réduction de ce type de bruit conduisent à des formules de paramétrisation de la soustraction spectrale classique ([1], [2]) en vue de rendre le débruitage plus flexible, mais sans pour autant surmonter le problème.

De nouvelles méthodes proposées incorporent des modèles psychoacoustiques pour modéliser les propriétés de notre système auditif afin de rendre moins audible et plus naturel ce type de bruit. Le phénomène de masquage simultané constitue le point clé de ces solutions. Il traduit la capacité d'un son puissant à masquer et rendre inaudible un autre plus faible se produisant simultanément. Ce phéomène est devenu largement exploité dans le débruitage de la parole de manière à masquer les composantes audibles du bruit et diminuer les distorsions du signal. Une modélisation des propriétés de masquage permet de calculer pour chaque trame du signal de parole une courbe de masquage représentant les points de pression acoustique nécessaires pour qu'un son soit audible en présence d'un masquant. Parmi les méthodes existantes, nous avons retenu celle de Johnston [3] pour sa simplicité de mise en oeuvre.

Dans ce papier, on s'intéresse à la réduction du bruit de type musical qu'engendrent des méthodes basées sur la soustraction spectrale et en particulier le filtrage de Wiener. On compare plusieurs méthodes qui introduisent des modifications perceptuelles du filtre de Wiener. Nous proposons aussi une nouvelle méthode qui améliore le signal débruité par le filtrage de Wiener en le pondérant par un second filtre perceptuel.

La section 2 est un rappel des notions de base sur le filtrage de Wiener standard et ses limitations. Dans la section suivante, on introduit les différentes méthodes perceptuelles avec lesquelles on va comparer la méthode que nous proposons dans ce même paragraphe. Avant de conclure, la section 4 présente les résultats expérimentaux issus de tests effectués sur la base de données de parole Tidigits, dans deux conditions de bruit et à différents rapport signal sur bruit.

2 Le filtrage standard et ses limitations

Soit un signal de parole bruité et échantillonné. Ce signal de parole bruité est divisé en trames successives qui comportent le même nombre d'échantillons noté N et qui se recouvrent de moitié. Le nombre N d'échantillons est choisi de manière à ce que la durée d'une trame soit de l'ordre de 30 ms.

Soient $y_k(t)$, $s_k(t)$ et $b_k(t)$, $t = 0, 1, \ldots, N-1$, le signal bruité, le signal propre et le bruit respectivement dans la $k^{\text{ème}}$ trame. On a donc, $y_k(t) = s_k(t) + b_k(t)$. Les Transformées de Fourier Discrètes (TFDs) de ces si-

gnaux sont respectivement notées $Y_k(\nu)$, $S_k(\nu)$ et $B_k(\nu)$, $\nu = 0, 1, \ldots, N-1$ et nous avons $Y_k(\nu) = S_k(\nu) + B_k(\nu)$. Le débruitage de la parole consiste à estimer les composantes fréquentielles $S_k(\nu)$ par un estimateur $H_k(\nu)$ tel que $\widehat{S}_k(\nu) = H_k(\nu)Y_k(\nu)$. L'erreur dûe à ce filtrage est la suivante :

$$\begin{aligned} e_k(\nu) &= \widehat{S}_k(\nu) - S_k(\nu) \\ &= (H_k(\nu) - 1)S_k(\nu) + H_k(\nu)B_k(\nu). \end{aligned} \quad (1)$$

L'expression $(H_k(\nu) - 1)S_k(\nu)$ représente la distorsion du signal tandis que $H_k(\nu)B_k(\nu)$ désigne le bruit résiduel. Ce dernier a un caractère musical dû à l'apparition de pics spectraux appelés tonales. Le filtrage de Wiener basé sur l'approche de Malah [4] est parmi les premières méthodes proposées pour pallier le problème du bruit musical. Certes, il parvient à réduire la quantité du bruit musical mais celui-ci reste toujours perceptuellement gênant. D'où l'intérêt d'introduire plus d'amélioration pour éliminer ce type d'artéfact.

L'expression du filtre de Wiener selon l'approche de Malah [4] est la suivante :

$$W_k(\nu) = \xi_k(\nu)/(1 + \xi_k(\nu)) \quad (2)$$

où

$$\xi_k(\nu) = (1-\alpha)h\left(\chi_k(\nu) - 1\right) + \alpha|\widetilde{S}_{k-1}(\nu)|^2/\gamma_k(\nu) \quad (3)$$

est l'estimation du Rapport Signal à Bruit (RSB) a priori

$$\mathbb{E}\left[|S_k(\nu)|^2\right] / \mathbb{E}\left[|B_k(\nu)|^2\right]. \quad (4)$$

Dans l'Eq. (3), $\widetilde{S}_{k-1}(\nu) = W_{k-1}(\nu)Y_{k-1}(\nu)$ est la ν^{th} composante spectrale du signal débruité par le filtrage de Wiener dans la trame $k-1$; $\gamma_k(\nu)$ est l'estimée de $\mathbb{E}\left[|B_k(\nu)|^2\right]$; $h(x) = x$ if $x \geq 0$ et $h(x) = 0$ ailleurs ; $\chi_k(\nu) = |Y_k(\nu)|^2/\gamma_k(\nu)$ est l'estimée du RSB a posteriori $|Y_k(\nu)|^2 / \mathbb{E}\left[|B_k(\nu)|^2\right]$; le coefficient α est fixé à 0.98 pour un meilleur compromis entre quantité du bruit musical et distorsion du signal [4].

L'estimée $\xi_k(\nu)$ prend en compte la trame courante avec un poids de $(1-\alpha)$ et la trame précédente avec un poids de α. L'effet de lissage dans l'estimation du RSB a priori dans l'Eq. (3) entraîne une diminution du niveau de bruit musical qui malgré tout reste présent et gênant à la perception.

3 Débruitage perceptuel : quelques approches récentes et une nouvelle méthode

Le diagramme ci-dessous résume toutes les méthodes décrites dans cette section et qui seront comparées dans la section suivante.

Ce synoptique permet d'améliorer le réducteur de bruit grâce à une pondération perceptuelle à travers un second filtrage $G_k(\nu)$. La courbe de masquage $T_k(\nu)$ est calculée à partir de l'estimée du signal propre obtenue à la sortie de Wiener. Quant à l'estimation $\gamma_k(\nu)$ de la densité spectrale du bruit, elle est mise au point pendant les instants de pause fournis par le Détecteur d'Activité Vocale (DAV) du standard ITU-T G729 (8kbits/s)[5].

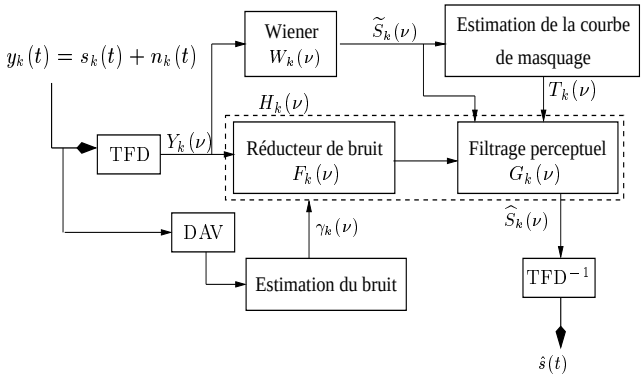

Fig. 1 – Diagramme général du débruitage proposé

Les méthodes que nous comparons peuvent toutes être décrites par la figure 1. Elles diffèrent par les expressions de $F_k(\nu)$ et de $G_k(\nu)$.

La première méthode (A) est un filtrage de Wiener de la quantité de bruit qui excède la courbe de masquage (le bruit audible) [6].

$$\text{(A)} : \begin{cases} F_k(\nu) = 1, \\ G_k(\nu) = \dfrac{|\widetilde{S}_k(\nu)|^2}{(|\widetilde{S}_k(\nu)|^2 + \max\left(\gamma_k(\nu) - T_k(\nu), 0\right))} \end{cases} \quad (5)$$

Quant à la deuxième méthode (B) introduite dans [7], le filtrage de Wiener, décrit à la section précédente, est activé uniquement pour les fréquences où le bruit est audible.

$$\text{(B)} : \begin{cases} F_k(\nu) = 1, \\ G_k(\nu) = \begin{cases} W_k(\nu) & \text{, si } \gamma_k(\nu) > T_k(\nu) \\ 1 & \text{, ailleurs.} \end{cases} \end{cases} \quad (6)$$

Remarque 3.1 L'objectif du débruitage perceptuel est de réduire le bruit sans apporter plus de distorsion sur le signal de parole. L'une des façons d'éviter des distorsions superflues est d'opérer uniquement dans les fréquences où le bruit est perceptuellement significatif. Cependant, en procédant ainsi, le bruit initialement inaudible, et par conséquent non pris en compte par le débruitage perceptuel, risque de devenir audible et gênant si les masquants de ce bruit sont filtrés.

La méthode que nous proposons (cf. Eq. (7)) vise à remédier à ce problème. Avec cette méthode, le filtrage de Wiener est activé même pour atténuer le bruit initalement inaudible mais il est ensuite pondéré par le filtre perceptuel G_k de l'Eq. (5).

$$\text{Double filtrage} : \begin{cases} F_k(\nu) = W_k(\nu), \\ G_k(\nu) = \dfrac{|\widetilde{S}_k(\nu)|^2}{(|\widetilde{S}_k(\nu)|^2 + \max\left(\gamma_k(\nu) - T_k(\nu), 0\right))} \end{cases} \quad (7)$$

Nous analysons maintenant les propriétés du double filtrage proposé dans l'Eq. (7) en utilisant le fait que $W_k(\nu)$ et $G_k(\nu)$ sont compris entre 0 et 1.

Si $\gamma_k(\nu) < T_k(\nu)$, ce qui signifie que le bruit est inaudible, nous avons $G_k(\nu) = 1$. Par contre le filtre de Wiener est toujours activé pour deux raisons : premièrement pour favoriser le gain en rapport signal sur bruit et deuxièmement pour réduire le risque que les portions de bruit inaudibles deviennent audibles lorsque les masquants ont disparu (voir la remarque 3.1).

Si $\gamma_k(\nu) \ll T_k(\nu)$, ce qui signifie un très bon RSB avant débruitage, on a $G_k(\nu) = W_k(\nu) = 1$. Aucune distorsion n'est alors introduite.

Si $\gamma_k(\nu) > T_k(\nu)$, on profite à la fois de la capacité du filtre de Wiener à réduire le bruit et de l'effet du facteur perceptuel pondérant pour traiter le bruit audible et améliorer la qualité du signal débruité par réduction du bruit musical.

Si $\gamma_k(\nu) \gg T_k(\nu)$, on a $\xi_k(\nu) \ll 1$. De fait, $W_k(\nu)G_k(\nu)$ tend plus rapidement vers 0 que $W_k(\nu)$. La méthode proposée accentue donc le débruitage quand le bruit est perceptuellement gênant.

Pour le double filtrage proposé et dans le but de réduire les discontinuités dans la fonction du gain $H_k = W_k G_k$ dûes au traitement sélectif en fréquence, on propose d'appliquer au filtre H_k un lissage fréquentiel par corrélogramme lissé, avant de procéder au débruitage. Ce lissage est le résultat de la convolution circulaire entre la suite de valeurs $H_k(\nu)$, $\nu = 0, 1, \ldots, N-1$, et une fenêtre de pondération dont les valeurs $C(\nu)$ pour $\nu = 0, 1, \ldots, N-1$, sont réelles, sont telles que $C(N-\nu) = C(\nu)$ et vérifient la condition de normalisation $\sum_{k=0}^{N-1} C(\nu) = 1$. La fenêtre que l'on a choisi est la version normalisée d'une puissance de la fenêtre de Hanning, à savoir :

$$C(\nu) = \frac{(0.5 + 0.5\cos(2\pi\nu/N))^q}{\sum_{\nu=0}^{N-1}(0.5 + 0.5\cos(2\pi\nu/N))^q}, \qquad (8)$$

pour $\nu = 0, 1, \ldots, N-1$. Cette convolution s'obtient comme TFD du produit des TFDs inverses des suites $H_k(\nu)$ et $C(\nu)$, $\nu = 0, 1, \ldots, N-1$. L'effet du lissage est illustré par la figure 2 ci-dessous.

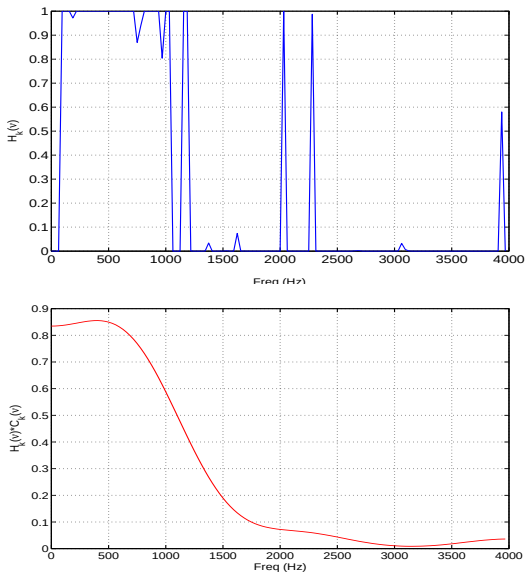

Fig. 2 – Effet lissage

Le dernier filtrage perceptuel ([9]) considéré (C) est conçu de façon à rendre le bruit musical inaudible en le forçant à être au dessous de la courbe de masquage, ce qui conduit à l'Eq. (9)

$$(\mathrm{C}) : \begin{cases} F_k(\nu) = 1, \\ G_k(\nu) = \min\left(\sqrt{T_k(\nu)/\gamma_k(\nu)}, 1\right) \end{cases} \qquad (9)$$

4 Résultats expérimentaux

L'étude expérimentale de la méthode proposée est menée sur des fichiers de parole de la base Tidigits sous-échantillonnés à 8 kHz. Les signaux de parole sont bruités par deux types de bruit additifs : un bruit blanc Gaussien et un bruit babble de la base NoiseX. On procède par trames de longueur $N = 256$, avec un recouvrement de 50%. La durée d'une trame est donc de 32 ms. Chaque trame est pondérée par la fenêtre de Hanning. On compare les cinq méthodes décrites ci-dessus, à savoir "A" (Eq. 5), "B" (Eq. 6), "C" (Eq. 9), le filtrage de Wiener standard et la méthode "Double filtrage" (Eq. 7). Le corrélogramme lissé utilisé dans le double filtrage est ajusté avec $q = 20$ dans l'Eq. (8).

Les performances de ces méthodes sont évaluées via deux critères objectifs : le rapport signal à bruit segmental SSNR (Segmental Signal to Noise Ratio) et un critère perceptuel, le MBSD (Modified Bark Spectral Distortion).

La figure 3 (resp. figure 4) présente les moyennes des notes MBSD et du SSNR pour 100 phrases de la base Tidigits bruités par un bruit blanc gaussien (resp. le bruit babble de la base NoiseX) à des RSBs en entrée de -5dB à 20dB. Les deux figures montrent bien l'apport de la méthode proposée par rapport aux autres méthodes testées. Le gain à la fois en SSNR et MBSD montre la capacité de la méthode à réduire le bruit et les distorsions.

5 Conclusion

D'après les mesures objectives de qualité présentées dans la section précédente, la méthode que nous proposons apporte une amélioration significative par rapport aux autres méthodes traitées dans ce papier. Nous envisageons des tests subjectifs et des tests de reconnaissance de la parole. Nous souhaitons utiliser le synoptique de la figure 1 pour tester d'autres algorithmes de débruitage perceptuel, tester d'autres DAV et d'autres techniques d'estimation du spectre de bruit, notamment dans la continuité de [10].

Références

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", Proc. of ICASSP 1979, Washington DC, 1979, pp. 208-211.

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. Speech and Audio Processing, vol. 7, pp. 126-137, 1999.

[3] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria", IEEE Jour. Selected Areas Commun., vol. 6, no. 2, pp. 314-323, 1988.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time

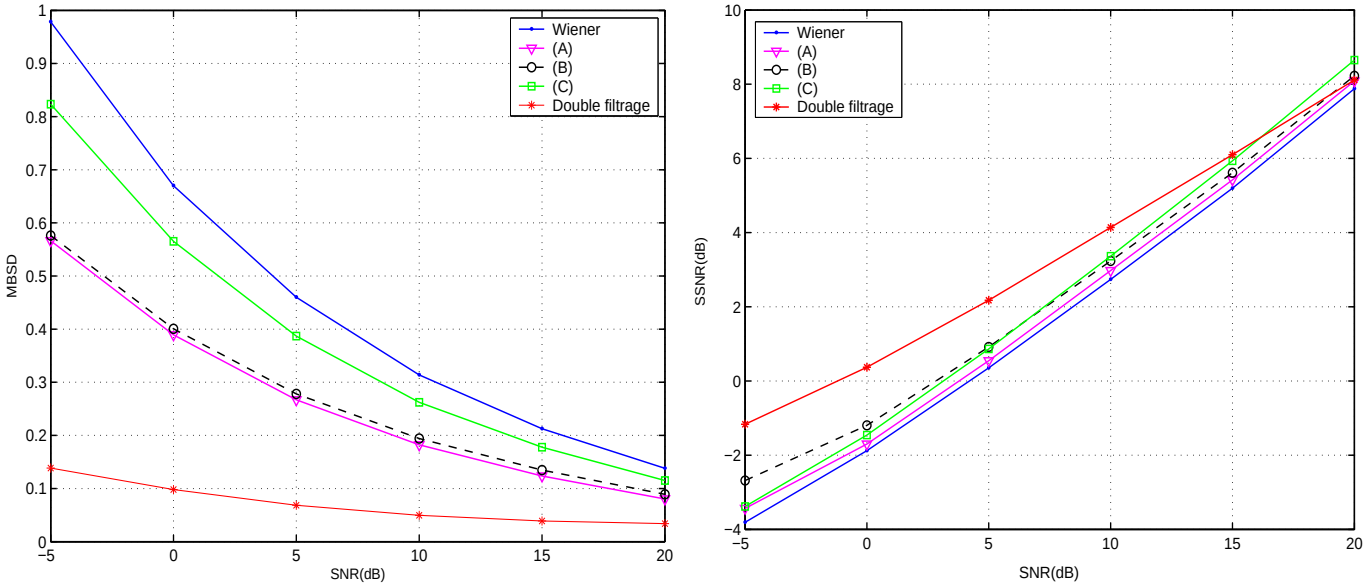

Fig. 3 – Comparaison des différentes méthodes de débruitage en terme de MBSD et de SSNR dans le cas de parole bruitée par un bruit blanc Gaussien

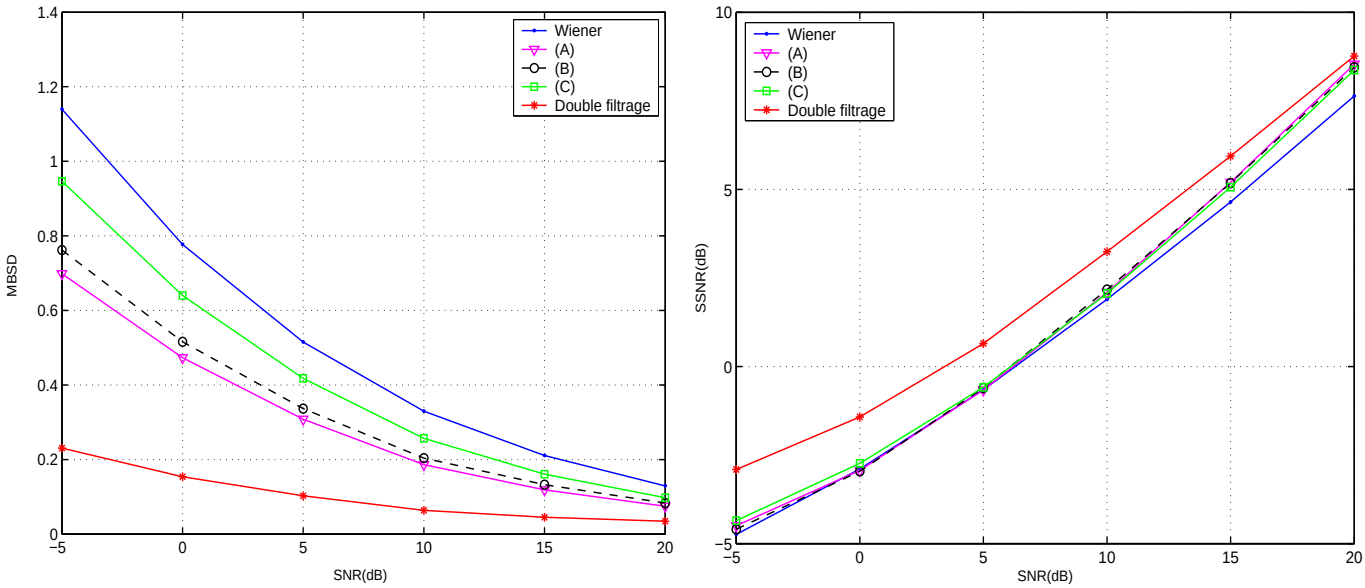

Fig. 4 – Comparaison des différentes méthodes de débruitage en terme de MBSD et de SSNR dans le cas de parole bruitée par du bruit babble de la base NoiseX

spectral amplitude estimator", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 1109-1121, 1984.

[5] IUT-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate structure algebraic-Code-Excited Linear Prediction (CS-ACELP)", 1996.

[6] L. Lin, W. H. Holmes and E. Ambikairajah, "Speech denoising using perceptual modification of Wiener filtering", IEEE Electronic Letters, vol. 38, no. 23, pp. 1486-1487, 2002.

[7] C. Beaugeant, V. Turbin, P. Scalart and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals", Signal Processing , Volume 64, Number 1, pp. 33-47(15), 1998.

[8] Y. Hu and P. Loizou, "Incorporating a psychoacoustic model in frequency domain speech enhancement", IEEE Signal Processing Letters, 11(2), pp. 270-273, 2004.

[9] T. Lee and Kaisheng Yao, "Speech enhancement by perceptual filter with sequential noise parameter estimation", Proc. of ICASSP 2004, Montreal , Quebec, Canada, pp. 693-696, 2004.

[10] A. Amehraye, D. Pastor, S. Ben Jebara, "On the Application of Recent Results in Statistical Decision and Estimation Theory to Perceptual Filtering of Noisy Speech Signals", Second International Symposium on Communications, Control and Signal Processing, ISCCSP 06, Marrakech, Maroc, 2006.