# Utilisation des Modèles de Markov Cachés pour le Débruitage

*Driss MATROUF et Jean-Luc GAUVAIN*

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
Tél.: 01 69 85 80 67 - Fax: 01 67 85 80 88
e-mail: driss@limsi.fr - http://www.lmsi.fr/TLP

## RÉSUMÉ

In this paper we address the problem of enhancing speech which has been degraded by additive noise. As proposed by Ephraim et al., autoregressive hidden Markov models (AR-HMM) for the clean speech and an autoregressive Gaussian for the noise are used. The filter applied to a given frame of noisy speech is estimated using the noise model and the autoregressive Gaussian having the highest *a posteriori* probability given the decoded state sequence. The success of this technique is highly dependent on accurate estimation of the best state sequence. A new strategy combining the use of cepstral-based HMMs, autoregressive HMMs, and a model combination technique, is proposed. The intelligibility of the enhanced speech is indirectly assessed via speech recognition, by comparing performance on noisy speech with compensated models to performance on the enhanced speech with clean-speech models. The results on enhanced speech are as good as our best results obtained with noise compensated models.

## 1. INTRODUCTION

Parmi les techniques de débruitage qui ont été les plus étudiées pendant les deux dernières décennies, on peut citer la technique de la soustraction spectrale [8, 9]. La soustraction spectrale, sous toutes ses formes, appliquée à un signal de parole bruité, afin de produire le signal propre correspondant, ne tire pas profit de l'information linguistique portée par ce signal. En fait, la plupart des implantations de la soustraction spectrale sont appliquées à un signal de parole exactement de la même manière qu'à n'importe quel autre signal bruité. Cette insuffisance d'information *a priori* dans le processus de débruitage engendre d'importantes distorsions dans les signaux restaurés.

Une tentative introduisant de l'information *a priori* dans le processus de débruitage a été proposée par J.S. Lim et A.V. Oppenheim [11]. Ils ont proposé une technique qui consiste à représenter le signal de parole par une succession de modèles AR : le signal est segmenté en une suite de zones stables (stationnaires), chaque zone est modélisée par un modèle AR. Les modèles AR et le signal propre sont tous les deux estimés en utilisant le signal bruité et la densité spectrale du bruit, supposée connue *a priori*. Pour cela, l'approche du MAP (*Maximum a Posteriori*) a été utilisée. La maximisation d'une fonction de vraisemblance appropriée est réalisée de manière itérative : 1) par rapport aux paramètres des modèles AR en supposant que le signal dont on dispose est propre, 2) par rapport au signal propre en utilisant les modèles AR et l'estimation de la densité spectrale du bruit. Il est évident que cette procédure itérative possède beaucoup plus de variables inconnues (signal propre et modèles AR) que de variables connues (signal bruité). Ceci conduit à des estimateurs, du signal et des modèles, avec de grandes variances. Pour pallier ce problème, Ephraim et al. [7] proposent d'estimer les modèles AR représentant le signal propre à partir de signaux d'apprentissage propres au lieu de les estimer directement sur le signal bruité.

Afin de déterminer les trames de parole propre maximisant la densité de probabilité *a posteriori*, Ephraim et al. [7] utilisent une procédure itérative fondée sur l'algorithme EM. Ce processus itératif est très dépendant de l'initialisation. Lorsque l'initialisation est mauvaise, ce processus converge vers un maximum local qui peut être très loin de la solution optimale. Cela est toujours le cas avec des signaux très bruités. Logan et Robinson [10] proposent d'utiliser une technique de combinaison de modèles dans le cadre des MMC autorégressifs (MMCAR) permettant de mieux initialiser ce processus. L'initialisation est obtenue par décodage du signal bruité en utilisant le système de reconnaissance fondé sur des MMCAR adaptés au bruit de test. Cependant, il est bien connu qu'un système de reconnaissance utilisant des MMC fondés sur le cepstre et ses dérivées est plus efficace qu'un système de reconnaissance utilisant des MMCAR, surtout dans le cas d'applications avec de grands vocabulaires. Dans cet article, nous proposons une extension de cette dernière approche en utilisant un système de reconnaissance fondé sur le cepstre et ses dérivées pour l'initialisation.

Nous proposons d'utiliser deux ensembles de modèles acoustiques : le premier est un MMC à densités continues (MMCDC) fondé sur le cepstre utilisé pour l'initialisation du processus itératif ; le second est un MMCAR utilisé pour estimer la succession de filtres à appliquer aux trames constituant le signal bruité. Pour une trame du signal bruité, nous cherchons la gaussienne cepstrale qui a la plus grande probabilité *a posteriori*. Cette recherche est faite par décodage du signal bruité en utilisant les modèles cepstraux adaptés au bruit de test. Le décodage fournit l'alignement trame/gaussienne-cepstrale, où chaque gaussienne cepstrale correspond à une gaussienne autoré-

gressive dans les MMCAR. Le filtre optimal est donc estimé en utilisant cette gaussienne autorégressive et la gaussienne autorégressive correspondant au bruit. Les MMC cepstraux et les MMCAR sont entraînés de façon à avoir une bijection entre les deux ensembles de modèles au niveau de la gaussienne : à une gaussienne dans les MMC cepstraux correspond une et une seule gaussienne dans les MMCAR. Pour cela, nous estimons en premier les MMC cepstraux et nous utilisons les statistiques correspondant à la dernière itération de l'algorithme EM pour estimer les paramètres des MMCAR.

Dans les sections suivantes nous décrivons la mise en œuvre de cette approche de débruitage. Des expériences avec différents types de bruits sont présentées afin de montrer l'apport de cette approche à l'amélioration de la qualité des signaux de parole bruités. Des expériences de reconnaissance avec les signaux débruités seront présentées .

## 2. LE PROCESSUS DE DÉBRUITAGE

Considérons, dans un premier temps, le cas d'une observation générée par une seule gaussienne. Soit $\mathbf{y}$ une trame du signal de parole bruité ($\mathbf{y} \in \mathbf{R}^K$) et $f_{\lambda_\mathbf{x}}$ la fonction de densité de probabilité (fdp) de la trame propre $\mathbf{x}$ correspondant à la trame bruitée $\mathbf{y}$. En supposant que $\mathbf{x}$ est générée par un processus autorégressif, sa fdp $f_{\lambda_\mathbf{x}}(\mathbf{x})$ est définie comme suit :

$$f_{\lambda_\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{K}{2}} |S_\mathbf{x}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}' S_\mathbf{x}^{-1} \mathbf{x})\} \quad (1)$$

où $S_\mathbf{x}$ désigne la matrice d'autocorrélation : $S_\mathbf{x} = \sigma_\mathbf{x}^2 (A_\mathbf{x}' A_\mathbf{x})^{-1}$, où $\sigma_\mathbf{x}^2$ est la variance de l'excitation du modèle AR, $A_\mathbf{x}$ est la matrice $K \times K$ triangulaire inférieure dont les $p+1$ premiers éléments de la première colonne sont les coefficients du processus AR : $a_{i,0 \leq i \leq p}$ où $a_0 = 1$.

Supposons également que la fonction de densité de probabilité du bruit additif peut être modélisée par une gaussienne autorégressive, soit $f_{\lambda_\mathbf{n}}$, le problème du débruitage consiste à estimer la trame propre $\mathbf{x}$ en utilisant $f_{\lambda_\mathbf{x}}$, $f_{\lambda_\mathbf{n}}$ et $\mathbf{y}$. Le critère du Maximum *A Posteriori* (MAP) peut être utilisé pour réaliser cette estimation :

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \log\{h(\mathbf{x}, \mathbf{y})\} \quad (2)$$

où $h(\mathbf{x}, \mathbf{y})$ désigne la fdp conjointe de $\mathbf{x}$ et $\mathbf{y}$. Le bruit étant additif et indépendant du signal, nous avons :

$$h(\mathbf{x}, \mathbf{y}) = f_{\lambda_\mathbf{x}}(\mathbf{x}) f_{\lambda_\mathbf{n}}(\mathbf{y} - \mathbf{x}). \quad (3)$$

Cette maximisation aboutit au filtre de Wiener bien connu dans la littérature du traitement de signal. La transformée de Fourier de l'estimation du signal propre est :

$$\hat{X}(\theta) = \frac{\Gamma_\mathbf{x}(\theta)}{\Gamma_\mathbf{x}(\theta) + \Gamma_\mathbf{n}(\theta)} Y(\theta) \quad (4)$$

où $Y(\theta)$ est la transformée de Fourier de la trame de parole bruitée, $\Gamma_\mathbf{x}(\theta)$ et $\Gamma_\mathbf{n}(\theta)$ sont les densités spectrales de puissance correspondantes aux deux processus AR $\mathbf{x}$ et $\mathbf{n}$. Les densités spectrales de puissance sont déterminées comme suit :

$$\Gamma_\mathbf{x}(\theta) = \frac{\sigma_\mathbf{x}^2}{|\Psi_\mathbf{x}(\theta)|^2}, \quad (5)$$

$$\Gamma_\mathbf{n}(\theta) = \frac{\sigma_\mathbf{n}^2}{|\Psi_\mathbf{n}(\theta)|^2} \quad (6)$$

où $\Psi_\mathbf{x}(\theta)$ et $\Psi_\mathbf{n}(\theta)$ désignent, respectivement, les transformées de Fourier des coefficients de prédiction associés à la trame de la parole propre et des coefficients de prédiction associés au bruit.

Considérons, maintenant, le cas général d'une observation générée par un MMCAR. Soit $\mathbf{y} = \mathbf{y}_{t,t=1,\ldots,T} / \mathbf{y}_t \in \mathbf{R}^k$ une séquence de trames correspondant à une phrase bruitée ($T$ est le nombre de trames constituant le signal correspondant à la phrase en question). L'estimateur MAP des trames correspondant au signal propre est obtenu itérativement en utilisant l'algorithme EM. À chaque itération $k$, la transformée de Fourier $\hat{X}_t^k$, $\{t = 1, \ldots, T\}$ de l'estimation de la trame propre $\hat{\mathbf{x}}_t(k)$ est obtenue par filtrage de la trame bruitée comme suit :

$$\hat{X}_t^{k+1}(\theta) = \left[ \sum_{\beta,\gamma} p_t(\beta, \gamma | \hat{\mathbf{x}}(k)) H_{\gamma|\beta}^{-1} \right]^{-1} Y_t(\theta), \quad (7)$$

où $p_t(\beta, \gamma | \hat{\mathbf{x}}(k))$ est la probabilité que la trame $\hat{\mathbf{x}}_t(k)$ soit générée à l'instant $t$ par la gaussienne $\gamma$ de l'état $\beta$, sachant que $\hat{\mathbf{x}}(k)$ est générée par le MMCAR. $H_{\gamma|\beta}$ est le filtre de Wiener associé à la gaussienne autorégressive $\gamma$ de l'état $\beta$ et à la gaussienne autorégressive du bruit (Equation 4).

Le succès de cette procédure d'estimation est très dépendant de la qualité d'estimation des probabilités *a posteriori* $p_t(\beta, \gamma | \hat{\mathbf{x}}(k))$. Ces probabilités sont estimées en utilisant la procédure "avant-arrière". L'estimation de ces probabilités en utilisant les modèles acoustiques entraînés sur de la parole propre donne lieu à une procédure itérative convergeant vers un maximum local qui est loin d'être la solution optimale recherchée, surtout pour des rapport signal sur bruit (RSB) faibles. Pour obtenir une bonne estimation des probabilités *a posteriori*, nous décodons le signal bruité $\mathbf{y}$ en utilisant des modèles acoustiques obtenus par adaptation au bruit de test des modèles acoustiques cepstraux retenus pour la parole propre (MMCDC), c'est-à-dire les meilleurs modèles disponibles pour la parole bruitée.

Plusieurs techniques ont été proposées pour estimer les modèles représentant la parole bruitée à partir des modèles représentant la parole propre et du modèle du bruit. Parmi ces techniques, on peut citer la Combinaison Parallèle de Modèles (CPM) et l'approche par génération de données [2]. Dans ce travail, nous utilisons une technique fondée sur la composition directe sur les données (CPD : Combinaison Parallèle de Données) [1]. Elle consiste à réutiliser pour chaque gaussienne d'un état donné, d'un modèle donné, les vecteurs cepstraux d'apprentissage qui ont permis son estimation. Comme pour la CPM, on suppose que le bruit n'affecte pas la distribution des trames entre les gaussiennes d'un état donné. Afin d'améliorer d'avantage l'estimation des probabilités *a posteriori*, il est possible, en plus de la CPD, d'utiliser des techniques d'adaptation aveugles comme la MLLR [5]. Dans ce travail, seule la CPD est utilisée pour l'adaptation, les composantes principales du système de débruitage proposé dans cet article sont présentées dans la Figure 1.

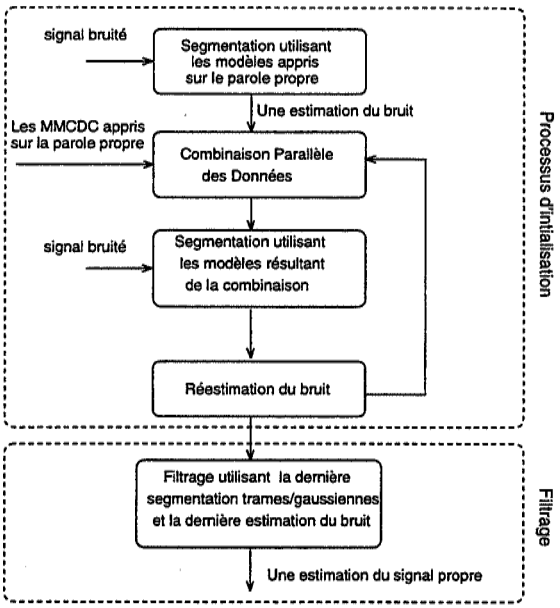


**Figure 1:** Le processus de débruitage utilisant le filtrage de Wiener dépendant de l'état dans les MMCAR. Les MM-CDC cepstraux et la CPD sont utilisés pour avoir une bonne estimation des probabilités *a poteriori*.

## 3. RÉSULTATS EXPÉRIMENTAUX

Afin d'évaluer cette technique de débruitage, nous utilisons un système entraîné sur 22148 phrases prononcées par 460 locuteurs, provenant du corpus MASK [4]. Les phrases prononcées par 450 locuteurs ont été utilisées pour l'apprentissage du système et les phrases prononcées par les 10 locuteurs restants ont été utilisées pour le test. Ces données ont été enregistrées avec un microphone de tête résultant en un RSB d'environ 35dB. Le signal de parole a subi un filtrage passe-bas avec une fréquence de coupure égale à 8kHz et a été échantillonné à 16kHz.

Pour les deux systèmes (cepstral et autorégressif) les vecteurs paramétriques sont estimés tous les 10ms sur une fenêtre de 30ms. Pour le système cepstral, les vecteurs paramétriques caractérisant une trame sont composés de 13 coefficients cepstraux auxquels sont ajoutées leurs dérivées premières et secondes (vecteur à 39 paramètres). La soustraction du cepstre moyen est réalisée sur chaque phrase. L'ordre des gaussiennes autorégressives est fixé à 16 pour les MMCAR et pour le modèle autorégressif correspondant au bruit. 608 phonèmes dépendant du contexte sont modélisés, chaque phonème dépendant du contexte est un MMCDC à trois états avec un mélange de gaussiennes comme densité d'observation (typiquement 20 gaussiennes).

L'algorithme de débruitage a été appliqué à des signaux de parole dégradés par un bruit additif. Dans cet article trois bruits extraits de la base de données NOISEX-92 [12] ont été utilisés : le bruit blanc, le bruit d'hélicoptère (Lynx) et le bruit d'avion à réaction (F16 Jet). Le signal restauré est obtenu en utilisant la technique standard d'*overlap-add* sur les segments résultant du filtrage, suivi du fenêtrage par Hanning.

**Amélioration de la qualité des signaux de parole bruitée**

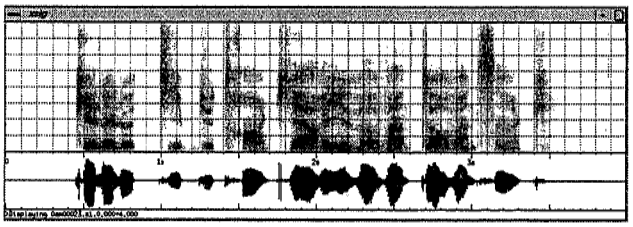

**Figure 2 :** *Spectrogramme du signal propre : "quel est le type du train qui arrive à 20 heures 25."*

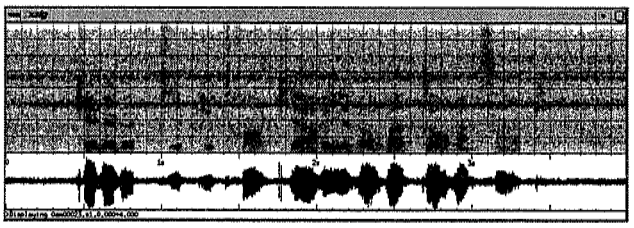

**Figure 3 :** *Signal bruité (RSB=5.7dB) généré en additionnant le bruit F16 Jet au signal de la Figure 2.*

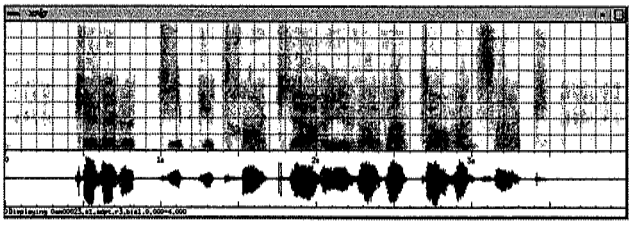

**Figure 4 :** *Signal obtenu par débruitage du signal de la Figure Figure 3. Le débruitage utilise le filtrage de Wiener dépendant de l'état.*

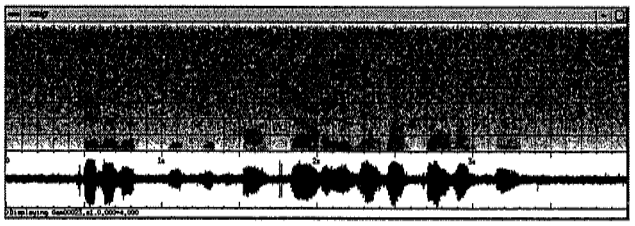

**Figure 5 :** *Signal bruité (RSB=1dB) généré en additionnant le bruit blanc au signal de la Figure 2.*

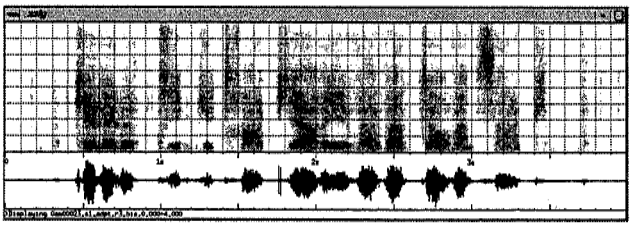

**Figure 6 :** *Signal obtenu par débruitage du signal de la Figure Figure 5. Le débruitage utilise le filtrage de Wiener dépendant de l'état.*

Dans la Figure 2 nous présentons le spectrogramme d'un signal propre. Les versions bruitées de ce signal sont obtenues en additionnant le bruit d'avion à réaction (F16 Jet) (Figure 3) et le bruit blanc (Figure 5). Les spectrogrammes des signaux obtenus par débruitage par application du filtrage de Wiener dépendant de l'état sont présentés dans les Figures 4 et 6 respectivement. Visuellement, une diminution considérable du bruit est observée par débruitage. En écoutant les signaux restaurés nous constatons une amélioration significative de la qualité avec de très faibles distorsions et l'absence du bruit musical.

**La reconnaissance avec les signaux débruités**

Afin d'évaluer objectivement cette procédure de débruitage, des expériences de reconnaissance avec des signaux débruités ont été réalisées. Ces expériences sont présentées dans le but de montrer qu'il n'y a pas de perte d'intelligibilité du point de vue du système de reconnaissance et non dans le but d'améliorer le taux de reconnaissance. En effet, la théorie prédit qu'aucun gain significatif en terme de taux de reconnaissance n'est à espérer par débruitage par rapport aux techniques de bruitage si les modèles mis en jeu sont de même qualité. L'objectif est l'amélioration de la qualité du signal sans perte d'intelligibilité.

Les taux d'erreur sur les mots dans différentes configurations sont présentés dans la table 1. En plus du filtrage de Wiener donné par l'équation 4 (filtre 1), nous avons utilisé le filtre résultant de l'application de la racine carrée au filtre de Wiener (filtre 2).

| | *% d'erreur sur les mots* | | |
|---|---|---|---|
| *Configuration de test* | F16 | Lynx | Blanc |
| Propre | 5.9 | 5.9 | 5.9 |
| Bruitée | 55.4 | 60.7 | 79.9 |
| Compensation (CPD) | 13.6 | 21.4 | 15.2 |
| Débruitage (filtre 1) | 13.9 | 21.7 | 15.2 |
| Débruitage (filtre 2) | 12.2 | 20.3 | 14.5 |

**Table 1** *Taux d'erreur moyens sur les mots dans différentes configurations de test. La colonne 1 correspond au bruit de F16 Jet (RSB=6.4), la colonne 2 correspond au bruit de Lynx (RSB=5.5dB) et la colonne 3 correspond au bruit blanc (RSB=1dB).*

La table 1 montre une très grande détérioration des performances lorsque le système est entraîné sur des signaux propres et testé sur des signaux bruités. Par exemple, lorsque le système est entraîné sur des signaux propres et testé sur des signaux bruités (bruit blanc : RSB=1dB), le taux d'erreur sur les mots augmente de 5.9% à 79.9%. La compensation du bruit de test en utilisant la CPD apporte un gain significatif (comparer la ligne 2 et 3). En utilisant des modèles acoustiques entraînés sur des signaux propres pour décoder les signaux restaurés par filtrage de Wiener dépendant de l'état, nous obtenons des résultats similaires à ceux obtenus en utilisant la CPD. En utilisant le système entraînés sur des signaux propres pour décoder les signaux restaurés par filtrage utilisant le filtre 2, un gain relatif de 4.6% est observé dans le cas du bruit blanc, 10.3% dans le cas du bruit F16 Jet, et 5.1% dans le cas du bruit de Lynx, par rapport aux résultats obtenus en utilisant la CPD. Ces résultats montrent que du point de vue du système de reconnaissance, la technique de débruitage proposée ne provoque pas de perte d'intelligibilité des signaux restaurés par rapport aux signaux bruités.

## 4. CONCLUSION

Dans cet article nous avons étudié et expérimenté une approche de débruitage fondée sur le critère du maximum *a posteriori* utilisant des MMC autorégressifs. Cette technique introduite par Ephraim et al. utilise l'algorithme EM qui nécessite dans ce cas une bonne initialisation pour fonctionner correctement. Nous avons proposé une nouvelle stratégie d'initialisation combinant l'utilisation de MMCAR, l'utilisation de MMCDC fondés sur le cepstre et ses dérivées et l'utilisation d'une technique de compensation de bruits que nous avons présenté dans des travaux ultérieurs [6]. Cette stratégie d'initialisation rend ce processus de débruitage très efficace même avec des rapports signal sur bruit faibles. Les expériences ont été réalisées en utilisant trois types de bruits provenant de la base de données NOISEX. Nous avons constaté une augmentation considérable de la qualité des signaux débruités, avec peu de distorsion et sans bruit musical. Des tests de reconnaissance avec les signaux débruités montrent qu'il n'y a pas de diminution de l'intelligibilité du signal débruité par rapport au signal bruité du point de vue du système de reconnaissance.

## REFERENCES


[1] J.L. Gauvain, L. Lamel, G.Adda, D.Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task," *ICASSP-96*.

[2] M.J.F. Gales, S.J. Young, "A fast and flexible implementation of parallel model combination", *ICASSP-95*, pp. 133-136.

[3] A. Acero, R.M. Stern, "Environmental Robustness in Automatic Speech Recognition," *IEEE Acoustics, Speech & Signal Processing*, pp. 849-852. April 1990.

[4] L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information," *Eurospeech'95*, Madrid.

[5] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech & Language*, pp. 171-185, 1995.

[6] D.Matrouf, J.L. Gauvain, "Model compensation for noises in test and training data," *ICASSP-97*.

[7] Y. Ephraim, D. Malah, B.H Juang, "On the Application of Hidden Markov Models for Enhancing Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37. NO. 12. December 1989.

[8] S.F Boll, " Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE*, pp. 113-120, 1979.

[9] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", *ICASSP-79*, pp. 208-211.

[10] B. T. Logan, A. J. Robinson, "Enhancement and Recognition of Noisy Speech within an AutoRegressive HMM Framework Using Noise Estimates from The Noisy Signal," *ICASSP-97*.

[11] J.S. Lim, A.V. Oppeenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 26, pp. 197-210, June 1978.

[12] A.P. Varga, H.J.M Steeneken, M. Tomlinson, D. Jones,"The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *In Technical Report, DRA Speech Research Unit*, 1992.