Utilisation des modèles de Markov cachés pour le débruitage

1. introduction

L'objectif de ce travail est l'amélioration de la qualité de signaux de parole bruitée tout en minimisant la perte d'intelligibilité pouvant être causée par les traitements effectués sur ces signaux.

La soustraction spectrale [1, 2, 13, 9] est l'une des techniques de débruitage qui a été la plus étudiée. Cette technique appliquée à un signal de parole bruité ne tire pas profit de l'information linguistique portée par ce signal. En fait, la soustraction spectrale est souvent appliquée à un signal de parole exactement de la même manière qu'à n'importe quel autre signal. Ce manque d'information *a priori* dans le processus de débruitage engendre d'importantes distorsions dans les signaux restaurés.

L'utilisation d'information *a priori* sur la nature du signal a été proposée par J.S. Lim et A.V. Oppenheim [14]. Cette approche consiste à représenter le signal de parole par une succession de modèles autorégressifs (AR) : le signal est découpé en une suite de segments de même durée où chaque segment est modélisé par un modèle autorégressif. La méthode proposée est fondée sur le critère du *Maximum a Posteriori* (MAP) : le signal débruité et les modèles AR sont déterminés en maximisant leur probabilité conjointe connaissant le signal bruité et une estimation de la densité spectrale du bruit. Cette maximisation est réalisée de manière itérative : 1) par rapport aux paramètres des modèles AR en supposant que le signal est non bruité, 2) par rapport au signal non bruité en utilisant les modèles AR et une estimation de la densité spectrale du bruit. Cette procédure ayant beaucoup plus de variables inconnues (signal propre et modèles AR) que de variables connues (signal bruité), elle conduit en général à une mauvaise estimation du signal et des modèles.

Pour pallier ce problème, Ephraim *et al.* [3] ont proposé d'utiliser des modèles de Markov cachés autorégressifs (MMC-AR) estimés sur des signaux non bruités au lieu de modèles AR estimés directement sur le signal bruité. Afin de déterminer les trames de la parole propre Ephraim *et al.* [3] utilisent une procédure itérative fondée sur l'algorithme EM (*Expectation – Maximization*). Ce processus itératif, qui est très dépendant de l'initialisation, converge vers un maximum local qui peut être très éloigné de la solution optimale, en particulier pour des signaux très bruités. Logan et Robinson [12] ont proposé d'utiliser une technique de combinaison de modèles pour MMC-AR afin de mieux initialiser ce processus. L'initialisation est alors effectuée en décodant le signal bruité au moyen de MMC-AR adaptés au bruit.

Le processus de débruitage proposé est fondamentalement dépendant de l'estimation de la probabilité qu'à un instant donné une gaussienne du MMC-AR ait généré la trame de parole propre non observée étant donné le signal bruité. Dans la suite de l'article ces probabilités sont appelées « probabilités *a posteriori* ». Elles peuvent être obtenues par décodage du signal de parole au moyen des modèles MMC-AR, mais il est un fait que pour reconnaître la parole, les MMC cepstraux sont significativement plus efficaces que les MMC-AR. C'est la raison pour laquelle nous proposons une nouvelle stratégie combinant l'utilisation de MMC autorégressifs et de MMC cepstraux associés à une technique d'adaptation des modèles au bruit.

L'article est organisé comme suit. Le principe du débruitage est décrit dans la section 2. L'approche proposée pour le calcul des probabilités *a posteriori* fait l'objet de la section 3 et le processus d'adaptation des MMC cepstraux au bruit est décrit en section 4. Enfin dans la section 5, nous présentons le fonctionnement global du système ainsi que les résultats expérimentaux.

2. le processus de débruitage

Le processus de débruitage est tout d'abord décrit pour une trame de parole supposée générée par une gaussienne AR, puis pour une suite de trames générées par un MMC-AR.

2.1. cas d'une gaussienne AR

Considérons, dans un premier temps, le cas d'une observation générée par une seule gaussienne AR. Soit $\mathbf{y}$ une trame du signal de parole bruité (typiquement 30 ms de signal, $\mathbf{y} \in \mathbf{R}^K$) et $f_{\lambda_\mathbf{x}}$ la densité de probabilité (ddp) du signal non bruité $\mathbf{x}$ correspondant à la trame bruitée $\mathbf{y}$. En supposant que $\mathbf{x}$ est générée par un processus autorégressif, sa ddp $f_{\lambda_\mathbf{x}}(\mathbf{x})$ 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 $S_\mathbf{x}$ est la matrice d'autocorrélation $S_\mathbf{x} = \sigma_\mathbf{x}^2 (A_\mathbf{x}' A_\mathbf{x})^{-1}$, $\sigma_\mathbf{x}^2$ est la variance

du résiduel du modèle AR, et $A_{\mathbf{x}}$ est une matrice de Toeplitz $K \times K$ triangulaire inférieure dont les $p+1$ premiers éléments de la première colonne sont les coefficients du processus AR : $a_{i,0 \leqslant i \leqslant p}$ avec $a_0 = 1$.

Il est supposé également que la densité de probabilité $f_{\lambda_{\mathbf{n}}}$ du bruit est une gaussienne autorégressive. Le problème consiste alors à estimer la trame débruitée en utilisant $f_{\lambda_{\mathbf{x}}}$, $f_{\lambda_{\mathbf{n}}}$ et $\mathbf{y}$. Le critère du *maximum a posteriori* peut être utilisé pour réaliser cette estimation :

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \log h(\mathbf{x}, \mathbf{y}) \qquad (2)$$

où $h(\mathbf{x}, \mathbf{y})$ est la dpp conjointe de $\mathbf{x}$ et $\mathbf{y}$. Le bruit étant additif et indépendant du signal, on a la relation :

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

Cette solution correspond au filtre de Wiener bien connu dans la littérature du traitement de signal. La transformée de Fourier du signal débruité $\hat{\mathbf{x}}$ est

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

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

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

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

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 non bruitée et des coefficients de prédiction associés au bruit.

2.2. cas d'un MMC-AR

Un MMC-AR est un modèle markovien caché [16] dont les densités de probabilité associées aux états sont des mélanges de gaussiennes autorégressives. Soit $f_{\lambda_{\mathbf{x}}}$ la densité de probabilité associée à un MMC-AR d'une séquence d'observation $\mathbf{x}$, où $\lambda_{\mathbf{x}}$ désigne les paramètres du modèle. On suppose que le MMC-AR est formé de N états et que à chaque état est associé un mélange de M gaussiennes. Soit $\mathbf{x} = x_{t,t=0,\ldots,T}$ une séquence de vecteurs générés par le modèle, $\mathbf{e} = e_{t,t=0,\ldots,T}$ la succession d'états qui a généré l'observation $\mathbf{x}$ (x_t a été généré par e_t), $\mathbf{h} = h_{t,t=0,\ldots,T}$ la succession de gaussiennes AR qui a généré $\mathbf{x}$ (x_t a été généré par h_t), la densité de probabilité de $\mathbf{x}$ est obtenue comme suit :

$$f_{\lambda_{\mathbf{x}}}(\mathbf{x}) = \sum_{\mathbf{e}} \sum_{\mathbf{h}} p_{\lambda_{\mathbf{x}}}(\mathbf{e}) p_{\lambda_{\mathbf{x}}}(\mathbf{h}|\mathbf{e}) g_{\lambda_{\mathbf{x}}}(\mathbf{x}|\mathbf{h}, \mathbf{e}) \quad (7)$$

où $p_{\lambda_{\mathbf{x}}}(\mathbf{e})$ désigne la probabilité d'observer la séquence d'états $\mathbf{e}$, $p_{\lambda_{\mathbf{x}}}(\mathbf{h}|\mathbf{e})$ est la probabilité d'observer la séquence de gaussiennes $\mathbf{h}$ sachant que la séquence d'états observés est $\mathbf{e}$, et $g_{\lambda_{\mathbf{x}}}(\mathbf{x}|\mathbf{e}, \mathbf{h})$ est la densité de probabilité de $\mathbf{x}$ connaissant $\mathbf{e}$ et $\mathbf{h}$. Cette dernière peut s'écrire :

$$g_{\lambda_{\mathbf{x}}}(\mathbf{x}|\mathbf{h}, \mathbf{e}) = \prod_{t=0}^{T} b_{\lambda_{\mathbf{x}}}(x_t|h_t, e_t) \quad (8)$$

où chaque terme du produit a la forme suivante :

$$b_{\lambda_{\mathbf{x}}}(x_t|h_t = \gamma, e_t = \beta) = \frac{1}{(2\pi)^{\frac{K}{2}} \left|S_{\gamma|\beta}\right|^{\frac{1}{2}}} exp\{-\frac{1}{2}(x_t^{'} S_{\gamma|\beta}^{-1} x_t)\} \quad (9)$$

avec $S_{\gamma|\beta} = \sigma_{\gamma|\beta}^2 (A_{\gamma|\beta}^{'} A_{\gamma|\beta})^{-1}$, où $\sigma_{\gamma|\beta}^2$ désigne l'énergie du résiduel. $A_{\gamma|\beta}$ est une matrice $K \times K$ de Toeplitz triangulaire, dans laquelle les $p+1$ premiers éléments de la première colonne constituent les coefficients du processus AR :

$$a_{\gamma|\beta} = (a_{\gamma|\beta}^0, a_{\gamma|\beta}^1, \ldots, a_{\gamma|\beta}^p)$$

et $a_{\gamma|\beta}^0 = 1$. La première étape du processus de débruitage consiste à estimer les paramètres du MMC-AR modélisant la parole propre en utilisant l'algorithme de réestimation de Baum [16].

Ephraim [3] utilise l'algorithme EM pour réaliser une estimation de type MAP du le signal non bruité à partir du signal bruité et des connaissances *a priori* sur le bruit. Etant donnée $\mathbf{y}$ une séquence de $T+1$ vecteurs représentant une séquence de segments de parole bruitée, l'approche du MAP consiste à déterminer le signal $\hat{\mathbf{x}}$ qui maximise $p(\hat{\mathbf{x}}|\mathbf{y})$. Ceci est réalisé au moyen de l'algorithme EM : soit $\hat{\mathbf{x}}(k)$, une estimation courante du signal non bruité (la première estimation est le signal bruité), une nouvelle estimation $\hat{\mathbf{x}}(k+1)$ vérifiant

$$p(\hat{\mathbf{x}}(k+1)|\mathbf{y}) \geqslant p(\hat{\mathbf{x}}(k)|\mathbf{y})$$

est obtenue en maximisant la fonction auxiliaire suivante :

$$\mathbf{\Phi}(\hat{\mathbf{x}}(k+1)) = \sum_{\mathbf{e},\mathbf{h}} p_{\lambda_{\mathbf{x}}}(\mathbf{e}, \mathbf{h}|\hat{\mathbf{x}}(k)) \log\left(p_{\lambda_{\mathbf{x}}}(\mathbf{e}, \mathbf{h}, \hat{\mathbf{x}}(k+1)|\mathbf{y})\right). \quad (10)$$

Cette fonction est directement maximiser en dérivant par rapport à $\hat{\mathbf{x}}(k+1)$ et en égalisant à zéro. On obtient, dans le domaine spectral, la relation suivante :

$$\widehat{X}_t^{k+1}(\theta) = \left[\sum_{\beta,\gamma} p_t(\beta, \gamma|\hat{\mathbf{x}}(k)) H_{\gamma|\beta}^{-1}\right]^{-1} Y_t(\theta), \quad (11)$$

où $p_t(\beta, \gamma|\hat{\mathbf{x}}(k))$ est la probabilité que la trame à l'instant t soit générée par la gaussienne γ de l'état β. $H_{\gamma|\beta}$ est le filtre de Wiener associé à la gaussienne AR γ de l'état β et à la gaussienne AR modélisant le bruit (équation 4). L'équation 11 est très intuitive, le filtre appliqué à la trame x_t n'étant rien d'autre que la somme des filtres associés à toutes les gaussiennes du MMC, pondérés par les probabilités que ces gaussiennes aient généré la trame à l'instant t. Le calcul de ces probabilités *a posteriori* $p_t(\beta, \gamma|\hat{\mathbf{x}}(k))$ est l'objet de la section suivante.

3. estimation des probabilités *a posteriori*

Le succès de la procédure de débruitage fondée sur l'équation 11 est lié à la qualité de l'estimation des probabilités *a posteriori* $p_t(\beta, \gamma|\hat{\mathbf{x}}(k))$ qui est bien entendu

très dépendante des modèles utilisés : plus ces modèles sont riches d'informations acoustiques et linguistiques concernant le signal, meilleure est l'estimation de ces probabilités.

Dans ce travail nous utilisons un système de reconnaissance de la parole pour estimer ces probabilités. Les MMC cepstraux à distributions continues, qui représentent l'état de l'art en matière de systèmes de reconnaissance indépendants du locuteur, offrent des performances supérieures à celles obtenues avec des MMC-AR. Il y a deux raisons à cela : d'une part la matrice de covariance d'une gaussienne cepstrale permet de bien mieux représenter la variabilité de la parole qu'une gaussienne autorégressive dont les paramètres sont limités aux coefficients du processus AR, et d'autre part l'utilisation de paramètres différentiels ainsi que la normalisation du cepstre moyen rendent les modèles cepstraux plus robustes que les modèles autorégressifs.

L'approche adoptée dans ce travail combine ces deux types de modèles acoustiques. Le premier modèle est un MMC cepstral à distributions continues qui est utilisé pour décoder la parole et ainsi estimer les probabilité *a posteriori*, le second est un MMC-AR utilisé pour estimer la succession de filtres à appliquer au signal bruité. Pour chaque trame du signal bruité, on retient les gaussiennes cepstrales qui ont les plus grandes probabilités *a posteriori* en décodant le signal bruité au moyen de modèles cepstraux adaptés au bruit. Une approximation généralement utilisée est celle qui consiste à remplacer le filtre de l'équation 11 par le filtre associé à la gaussienne qui a la plus grande probabilité *a posteriori*. On considère dans ce cas qu'une trame est générée exclusivement par une seule gaussienne du MMC : l'association d'une trame à un état est déterminée en utilisant l'algorithme de Viterbi, et l'association de cette trame à une gaussienne est réalisée en choisissant la gaussienne ayant la plus grande probabilité *a posteriori*. Ce décodage produit un alignement trame/gaussienne-cepstrale, où chaque gaussienne cepstrale est associée à une gaussienne autorégressive du MMC-AR. Le filtre optimal est estimé en utilisant la séquence de gaussiennes autorégressives ainsi obtenue et la gaussienne autorégressive correspondant au bruit.

Le MMC cepstral et le MMC-AR sont construits de façon à assurer une correspondance entre les gaussiennes des deux modèles : à chaque gaussienne cesptrale est associée une et une seule gaussienne autorégressive. Cette correspondance est réalisée par construction : on procède tout d'abord à l'estimation des MMC cepstraux en utilisant l'algorithme EM ; les paramètres du MMC-AR sont ensuite estimés en utilisant les statistiques de la dernière itération dans le processus d'estimation du MMC cepstral. (Il est important de noter ici qu'une gaussienne cepstrale modélise un vecteur de coefficients cepstraux alors qu'une gaussienne autorégressive modélise directement le signal correspondant à un segment de parole).

On peut utiliser le MMC cepstral pour effectuer à la fois le décodage et le filtrage, mais dans ce cas il n'est plus possible d'intégrer les bandes critiques ni la normalisation cepstrale dans le calcul des coefficients cepstraux, ce qui peut réduire significativement la qualité de l'estimation des probabilités *a posteriori*.

Pour obtenir une bonne estimation des probabilités *a posteriori*, le signal bruité $\mathbf{y}$ est décodé en utilisant des modèles acoustiques obtenus par adaptation au bruit des modèles acoustiques cepstraux retenus pour la parole propre, c'est-à-dire les meilleurs modèles disponibles pour la parole bruitée (en ignorant l'effet Lombard). Plusieurs techniques ont été proposées pour estimer les modèles représentant la parole bruitée à partir des modèles représentant la parole propre et le modèle du bruit [5, 9]. Dans ce travail, une technique fondée sur la composition directe des données (CPD : combinaison parallèle de données) [6, 8] est utilisée. Dans la section suivante nous décrivons cette technique d'adaptation des modèles acoustiques au bruit.

4. la combinaison parallèle de données

Notons $\mathbf{x}$ le signal non bruité utilisé pour l'apprentissage, $\mathbf{n}$ les bruit supposé additif, et $\mathbf{y}$ le signal bruité, $\mathbf{y} = \mathbf{x} + \mathbf{n}$. Puisque les données d'apprentissage sont supposées non bruitées, les modèles cepstraux estimés sur ces données ne peuvent être utilisés pour décoder convenablement le signal bruité. Pour obtenir des modèles représentant le signal bruité, on utilise une technique inspirée de la composition parallèle de modèles (Parallel Model Combination, PMC) [7, 4], qui combine le modèle de parole bruitée à un modèle de bruit. Le modèle du bruit est généralement une gaussienne dont la moyenne et la matrice de covariance sont estimées à partir d'un échantillon représentatif du bruit.

Plusieurs techniques ont été proposées pour estimer les modèles correspondant à la parole bruitée. Parmi ces techniques, on peut citer l'approximation log-normale [4], l'intégration numérique, et l'approche par génération de données [5]. Dans le cadre de la PMC, l'approximation log-normale est souvent utilisée lors du passage du domaine spectral au domaine cepstral. Elle suppose que la somme de deux variables aléatoires log-normales est elle-même log-normale. Cette approximation devient inacceptable dans le cas de gaussiennes avec de grandes variances (par rapport à la variance globale de l'espace acoustique). En outre, la difficulté des calculs et le nombre d'approximations requises dans le cadre de la PMC augmentent rapidement avec la complexité des formules utilisées pour le calcul des paramètres cepstraux différentiels. La technique fondée sur la génération de données (Data Driven PMC - DDPMC) a été proposée par Gales et Young pour surmonter les problèmes liés à la PMC [5]. Elle consiste à générer aléatoirement les vecteurs associés à chaque état. Ces vecteurs sont ensuite combinés avec des trames générées aléatoirement à partir du modèle de bruit. Le résultat de la combinaison est utilisé pour réestimer la densité de probabilité associée à l'état en question en utili-

sant l'algorithme EM. Cette technique est nécessairement coûteuse en temps de calcul si on veut générer un nombre suffisant de vecteurs pour avoir une validité statistique.

Pour résoudre les problèmes liés à la PMC et à la DDPMC, la composition peut être directement effectuée sur les données [6, 8]. Cette technique de combinaison parallèle de données (CPD), nécessite l'identification pour chaque gaussienne de chaque état, des vecteurs cepstraux qui ont permis son estimation. La composition consiste alors à réestimer les paramètres de chaque gaussienne après avoir additionné dans le domaine spectral les trames correspondant à la gaussienne et des trames représentatives du bruit.

La combinaison des données d'apprentissage $\mathbf{x}$ avec l'estimation de $\mathbf{n}$ est réalisée phrase par phrase, ce qui permet de calculer les dérivées cepstrales premières et secondes de manière habituelle. Pour un vecteur donné associé à un état donné, l'association à une gaussienne est déterminée en choisissant celle qui a la plus grande probabilité *a posteriori*. Cette association entre les trames d'apprentissage et les gaussiennes des modèles est obtenue durant l'apprentissage du modèle sur les données non bruitées.

5. résultats expérimentaux

Le fonctionnement général du système de débruitage est présenté sur la figure 1. Le signal bruité est initialement décodé par un système de reconnaissance utilisant des modèles cepstraux estimés sur de la parole non bruitée. Ce décodage permet de segmenter le signal et d'obtenir une première estimation du bruit à partir des trames associées au modèle de silence. Cette estimation permet ensuite de dériver des modèles cepstraux de parole bruitée au moyen de la méthode CPD. Les modèles ainsi obtenus sont utilisés pour mieux segmenter le signal bruité et réestimer le bruit. Ces deux opérations peuvent être répétées autant de fois que

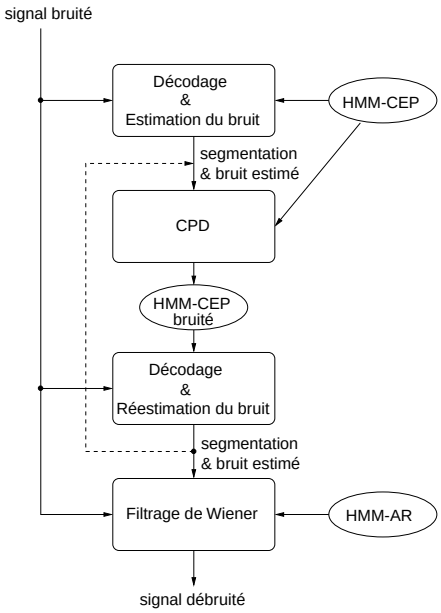

Figure 1. – Processus de débruitage utilisant le filtrage de Wiener avec des modèles de Markov cachés autorégressifs. Des modèles de Markov cepstraux sont utilisés pour obtenir une meilleure estimation des filtres.

nécessaire. Dans la pratique deux itérations suffisent pour converger vers un modèle correct. La dernière segmentation (algorithme de Viterbi) est utilisée pour estimer les filtres de Wiener à partir des gaussiennes AR du modèle de parole non bruité et de la gaussienne AR du bruit. Le signal débruité est obtenu par la technique standard de recouvrement et addition (*overlap-add*) sur les segments résultant du filtrage.

Afin d'évaluer cette technique de débruitage, nous avons utilisé un corpus d'environ 22000 phrases prononcées par 460 locuteurs (un sous-ensemble du corpus Mask [10]). Les phrases prononcées par 450 locuteurs ont été utilisées pour l'estimation des modèles acoustiques et du modèle de langage du système, et les énoncés des 10 locuteurs restants ont été utilisés pour les tests. Ces données ont été enregistrées dans un environnement non bruité avec un rapport signal sur bruit (RSB) d'environ 35dB. Le signal a été échantillonné à 16 kHz après un filtrage passe-bas à 8 kHz.

Pour les deux jeux de modèles (cepstral et autorégressif), les vecteurs paramétriques sont estimés toutes les 10ms sur des fenêtres de 30ms. Pour le modèle cepstral, le vecteur caractérisant une trame est composé de 13 coefficients cepstraux (MFCC) auxquels sont ajoutées leurs dérivées premières et secondes pour un total de 39 paramètres. La soustraction du cepstre moyen est effectuée phrase par phrase. Pour les gaussiennes autoregressives, l'ordre du modèle est fixé à 16.

Nous avons retenu un ensemble de 608 phonèmes en contexte (triphones), et chacune de ces unités est représentée par un MMC à trois états avec 20 gaussiennes par état. Le modèle de langage est un modèle bigramme estimé sur les transcriptions orthographiques des données d'apprentissage.

L'algorithme de débruitage a été appliqué à des signaux de parole dégradés par un bruit additif. Pour ces expériences, trois bruits extraits du corpus NOISEX-92 [17] ont été utilisés : un bruit blanc, un bruit d'hélicoptère (Lynx) et un bruit d'avion (F16).

Les techniques de débruitage sont souvent évaluées en estimant le RSB avant et après réduction du bruit. Les méthodes d'estimation du RSB ne prennent généralement pas en compte toutes les distorsions dues au débruitage, et ne permettent donc pas de mesurer la perte d'intelligibilité.

Pour évaluer notre procédure de débruitage, nous adoptons une approche qui consiste à mesurer, après réduction du bruit, le taux d'erreur d'un système de reconnaissance dont les paramètres des modèles ont été estimés sur de la parole non bruitée. Cette solution permet d'évaluer la perte d'intelligibilité du signal sans effectuer de tests de perception. Bien entendu, ce type d'évaluation ne peut pas remplacer des tests de perception dans la mesure où nous ne connaissons pas exactement la corrélation entre le taux de reconnaissance et la qualité (et l'intelligibilité) du signal. Néanmoins notre approche permet une évaluation à moindre coût et constitue un meilleur outil d'évaluation que la mesure du RSB.

Les expériences de reconnaissance sont présentées dans le but de montrer qu'il n'y

Tableau 1. – Taux d'erreur moyens sur les mots dans différentes configurations de test. La colonne 1 correspond au bruit de F16 (RSB = 6,4), la colonne 2 correspond au bruit de Lynx (RSB = 5,5dB) et la colonne 3 correspond au bruit blanc (RSB = 1dB). Les modèles acoustiques du système de reconnaissance sont estimés sur de la parole non bruitée.

Configuration de test	bruit de F16	bruit de Lynx	bruit Blanc
Parole non bruitée	5,9	5,9	5,9
Parole bruitée	55,4	60,7	79,9
Compensation (CPD)	13,6	21,4	15,2
Débruitage	13,9	21,7	15,2

a pas de perte d'intelligibilité du point de vue du décodeur linguistique et non dans le but d'améliorer le taux de reconnaissance. En effet, aucun gain significatif en terme de taux de reconnaissance n'est à espérer par débruitage par rapport aux techniques de bruitage si les critères à optimiser et les modèles mis en jeu sont de même nature. Les taux d'erreur sur les mots dans différentes configurations sont présentés dans le tableau 1.

Comme on pouvait si attendre, on observe une très grande détérioration des performances lorsque l'apprentissage est effectué sur des signaux non bruités et les tests sur des signaux bruités. Par exemple, le taux d'erreur augmente de 5,9 % à 79,9 % lorsque les données de test sont bruitées avec un bruit blanc correspondant à un RSB de 1dB (comparer les lignes 1 et 2 du tableau 1).

La compensation du bruit au moyen de la CPD apporte un gain très significatif (comparer les lignes 2 et 3 de le tableau 1). En utilisant des modèles acoustiques estimés sur des signaux non bruités 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, c'est-à-dire un gain relatif moyen (sur toutes les expériences) de l'ordre de 74 % par rapport au décodage du signal bruité en utilisant les modèles de parole non bruitée (comparer les lignes 2 et 4 du tableau 1). Ce qui montre que, du point de vue du système de reconnaissance, la technique de débruitage proposée n'engendre pas de perte d'intelligibilité par rapport aux signaux bruités.

De nombreux travaux sur la robustesse des systèmes de reconnaissance ont montré que les meilleurs résultats sont généralement obtenus par adaptation des modèles acoustiques au bruit, c'est-à-dire en bruitant les modèles. Les résultats dans le tableau 1 (comparaison des lignes 3 et 4) montrent que le débruitage du signal permet d'atteindre les mêmes performances que le bruitage des modèles à condition que les modèles et les critères mis en jeu soient de mêmes natures.

La figure 2a est le spectrogramme d'une phrase du corpus MASK. Les versions bruitées de ce signal sont obtenues en additionnant un bruit d'avion (F16) (figure 2b) et un bruit blanc (figure 2d). Les spectrogrammes des signaux obtenus par débruitage au moyen d'un filtrage de Wiener dépendant de l'état du MMC sont donnés sur les figures 2c et 2e. Une diminution très significative du bruit est observée après débruitage, le RSB dans les signaux débruités est de l'ordre de 35dB c'est-à-dire comparable au RSB du signal non bruité. En écoutant les signaux restaurés on peut constater une amélioration significative de la qualité avec de très faibles distorsions et l'absence de bruit musical.

6. conclusion

Dans cet article nous avons étudié une approche de débruitage fondée sur le critère du maximum *a posteriori* et des

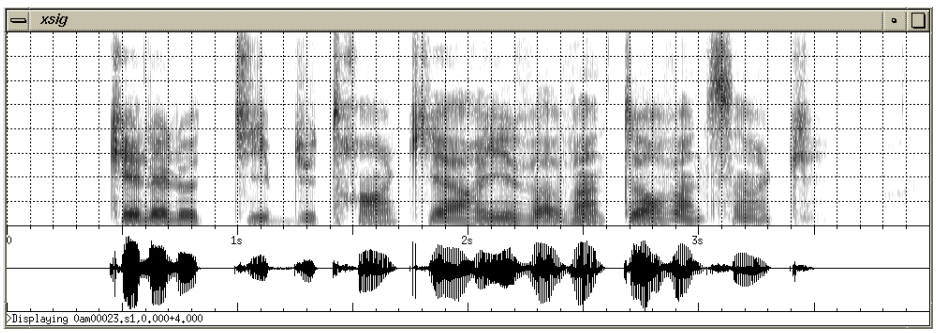

Figure 2a. – Spectrogramme du signal propre : « quel est le type du train qui arrive à 20 heures 25. »

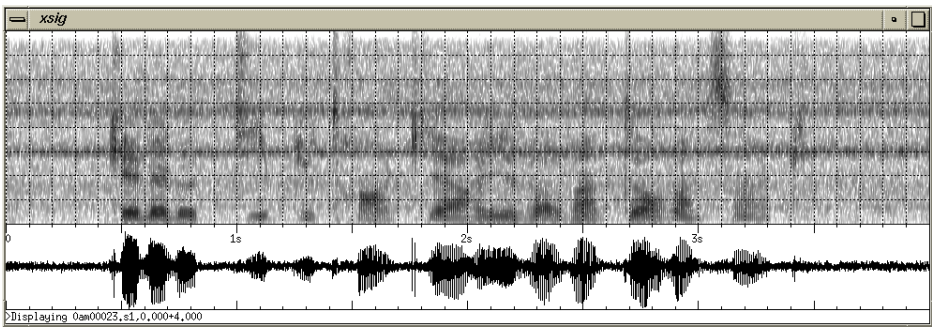

Figure 2b. – Signal bruité (RSB=5,7dB) généré en additionnant le bruit F16 Jet au signal de la Figure 2a.

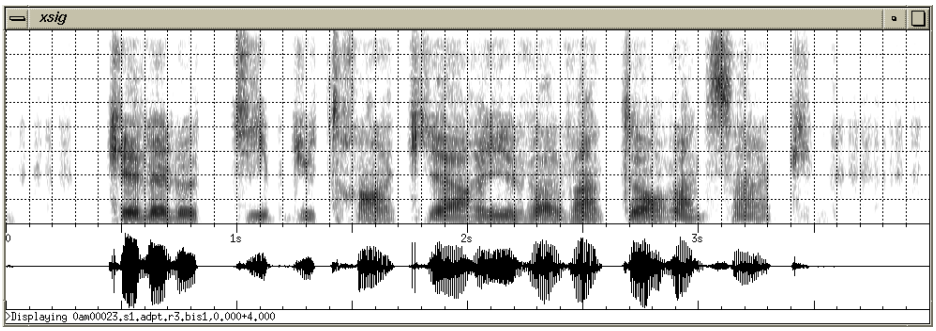

Figure 2c. – Signal obtenu par débruitage du signal de la Figure 2b. Le débruitage utilise le filtrage de Wiener dépendant de l'état.

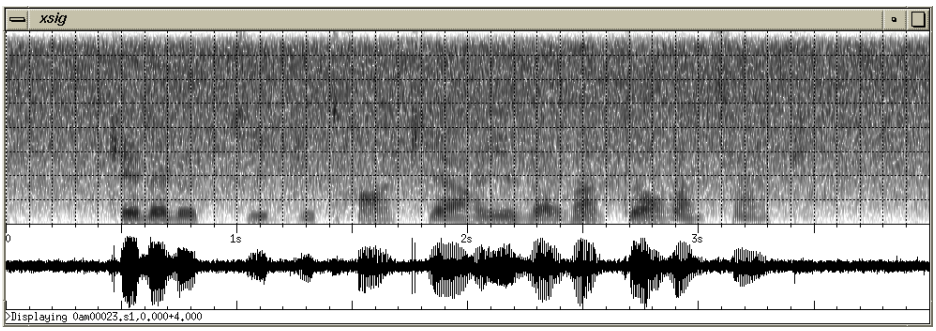

Figure 2d. – Signal bruité (RSB=1dB) généré en additionnant le bruit blanc au signal de la Figure 2c.

modèles de Markov cachés autorégressifs. Cette technique introduite par Ephraim *et al*. utilise l'algorithme EM qui ne converge vers une solution satisfaisante s'il n'est pas correctement initialisé. Nous avons proposé une méthode d'initialisation combinant l'utilisation de modèles autorégressifs et de modèles cepstraux associés à une technique de compensation de bruit par CPD [8].

Cette méthode rend le processus de débruitage très efficace même avec des rapports signal sur bruit faibles. Nous avons constaté une augmentation significative de la qualité des signaux débruités avec peu de distorsions et sans bruit musical. Les 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. Il est intéressant de noter, qu'en terme de reconnaissance, le débruitage permet d'atteindre les mêmes performances que l'adaptation des modèles acoustiques. Dans les deux cas, bruitage et débruitage, on observe un gain relatif moyen (toutes expériences confondues) de 74 % par rapport aux résultats obtenus par décodage du signal bruité en utilisant les modèles acoustiques non adaptés au bruit. Ce résultat n'est en fait pas surprenant puisque le processus de débruitage utilise les mêmes sources d'information que le système de reconnaissance, qu'il s'agisse du niveau acoustique ou linguistique.

Bien entendu, ce système de débruitage dépend de la nature de données puisqu'un système de reconnaissance est généralement spécifique à une application donnée. Cette dépendance qui concerne surtout le modèle linguistique peut être réduite en augmentant la couverture linguistique du système, ou plus simplement en n'effectuant qu'un décodage phonétique mais avec le risque de réduire l'intelligibilité du signal.

RÉFÉRENCES

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

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

[3] 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.

[4] M.J.F. Gales, S.J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech & Language*, 9(4), pp. 289-307, 1995.

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

[6] J.L. Gauvain, L. Lamel, G. Adda, D. Matrouf, "Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task", *ICASSP-96*, pp. 73-76.

[7] F. Martin, K. Shikano, Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models," *Eurospeech'93*, pp. 1031-1034.

[8] D. Matrouf, J.L. Gauvain, "Model compensation for noises in test and training data," *ICASSP-97*, pp. 831-834.

[9] C. E. Mokbel, G. F. A. Chollet, "Automatic Word Recognition in Cars," *IEEE Transactions on Speech and Audio Processing*, Vol. 3. No. 3. September 95.

[10] L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information," *Eurospeech'95*, pp. 1961-1964.

[11] 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.

[12] 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*, pp. 843-846.

[13] P. Lockwood, J. Boudy, M. Blanchet, "Non-linear Spectral Subtraction (NSS), and Hidden Markov Models for robust speech recognition in car noise environments," *ICASSP-92*, pp. 265-268.

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

[15] A.B. Poritz, "Linear predictive hidden Markov models and the speech signal", *ICASSP-82*, pp. 1291-1294.

[16] B.H. Juang, "Maximum-Liklihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains", *AT&T Technical Journal*, pp. 1235-1249, July-August 1985.

[17] 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.

LES AUTEURS

D. MATROUF, J.L. GAUVAIN, LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France, e-mail : {driss.gauvain}@limsi.fr