Statistiques d'ordre supérieur en vue de la détection de la parole

Arnaud Martin

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Les statistiques d'ordre supérieur, souvent négligées pour des raisons de difficultés de calculs, apportent des informations importantes. Les moments d'ordre 3 et 4 ont déjà été introduits dans des algorithmes de détection d'activité vocale. Ces travaux seront présentés au paragraphe 6. On se propose ici d'étudier, après une brève présentation d'éléments théoriques sur les moments et les cumulants, quelques estimateurs de ces statistiques. On tentera ensuite d'introduire ces informations dans l'algorithme de détection Bruit/Parole du CNET.

2 Quelques éléments théoriques sur les moments et les cumulants

Soit X une variable aléatoire réelle, de densité de probabilité $p_X(t)$. Sa fonction caractéristique $\phi_X(t)$, définie comme la transformée de Fourier de la densité de probabilité est donnée par :

$$\phi_X(t) = \int_{-\infty}^{+\infty} e^{itx} p_X(x) dx.$$

La fonction ϕ_X est continue, de module inférieur ou égal à 1, et $\phi_X(0) = 1$. Ainsi ϕ_X est non nulle dans un voisinage de l'origine, on peut donc définir son logarithme népérien pour t proche de l'origine :

$$\psi_X(t) = \ln(\phi_X(t)).$$

Cette fonction ψ_X est appelée *seconde fonction caractéristique*. *Les moments d'ordre* q sont définis par la dérivée d'ordre q à l'origine, de la première fonction caractéristique :

$$\mu_q = (-i)^q . \frac{d^q \phi_X(t)}{dt^q} \mid_{t=0} = E[X^q].$$

Les cumulants d'ordre q sont définis par la dérivée d'ordre q à l'origine, de la seconde fonction caractéristique :

$$\kappa_q = (-i)^q . \frac{d^q \psi_X(t)}{dt^q} \mid_{t=0} = Cum[X, X, ..., X].$$

Cette deuxième notation ne sera employée que lorsqu'il y a confusion possible. Par l'expression $\psi_X(t) = \ln(\phi_X(t))$, on peut exprimer les cumulants d'ordre q en fonction des moments d'ordre inférieur ou égal à q. On peut ainsi établir :

$$\kappa_1 = \mu_1,$$

$$\kappa_2 = \mu_2 - (\mu_1)^2,$$
$$\kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3,$$
$$\kappa_4 = \mu_4 - 4\mu_3\mu_1 - 3(\mu_2)^2 + 6\mu_2(\mu_1)^2 - (\mu_1)^4.$$

Remarquons que les cumulants d'ordre 1 et 2, sont respectivement la moyenne et la variance de la variable aléatoire réelle X. Les coefficients d'asymétrie (*skewness*) et d'aplatissement (*kurtosis*) sont les cumulants normalisés d'ordre 3 et 4 respectivement :

$$\chi_X = \frac{\kappa_3}{(\kappa_2)^{\frac{3}{2}}},$$
$$\gamma_X = \frac{\kappa_4}{(\kappa_2)^2}.$$

Ces coefficients permettent de préciser l'asymétrie et l'aplatissement de la distribution de la variable aléatoire X.

Dans le cas gaussien, on remarque que les cumulants d'ordre supérieur à deux sont tous nuls, et la seconde fonction caractéristique s'écrit uniquement, en fonction des moments d'ordre 1 et 2 :

$$\psi_X(t) = i\mu_1 t - \frac{1}{2}\mu_2 t^2.$$

Les variables aléatoires gaussiennes sont donc entièrement décrites par les moments du premier et second ordre. L'approximation gaussienne des distributions du bruit et de la parole, faite en considérant le théorème de la limite centrale, a restreint la recherche au second ordre.

Dans le cas des variables aléatoires multidimensionnelles, c'est-à-dire $X^T = [X_1, X_2, ..., X_{n-1}, X_n]$, les moments et les cumulants se déduisent des deux premières fonctions caractéristiques :

$$\phi_X(\mathbf{t}) = \int_{-\infty}^{+\infty} e^{i\mathbf{t}^T\mathbf{x}} p_X(\mathbf{x}) d\mathbf{x} = E[e^{i\mathbf{t}^T X}],$$
$$\psi_X(\mathbf{t}) = \ln(\phi_X(\mathbf{t})),$$

où $\mathbf{t}$ et $\mathbf{x}$ sont des vecteurs de dimension n.

On a ainsi :

$$\mu_{q_1,...,q_n} = (-i)^q \left(\frac{\delta}{\delta t_1}\right)^{q_1} \left(\frac{\delta}{\delta t_2}\right)^{q_2} ... \left(\frac{\delta}{\delta t_n}\right)^{q_n} \phi(\mathbf{t}) \mid_{\mathbf{t=0}},$$

et

$$\begin{aligned} \kappa_{q_1,...,q_n} &= (-i)^q \left(\frac{\delta}{\delta t_1}\right)^{q_1} \left(\frac{\delta}{\delta t_2}\right)^{q_2} ... \left(\frac{\delta}{\delta t_n}\right)^{q_n} \psi(\mathbf{t}) \mid_{\mathbf{t=0}} \\ &= Cum(X_1^{q_1}, X_2^{q_2}, ..., X_n^{q_n}), \end{aligned}$$

où $q = q_1 + q_2 + ... + q_n$, et $\mathbf{t}$ est un vecteur à n dimensions.

Si les variables aléatoires sont centrées, on a :

$$\kappa_{i,j} = \mu_{i,j},$$

$$\kappa_{i,j,k} = \mu_{i,j,k},$$

$$\kappa_{i,j,k,l} = \mu_{i,j,k,l} - \mu_{i,j}\mu_{k,l} - \mu_{i,k}\mu_{j,l} - \mu_{i,l}\mu_{j,k}.$$

À partir de la définition de la seconde fonction caractéristique, il est possible d'écrire, par la formule de Leonov et Shiryayev, les relations générales liant moments et cumulants :

$$\kappa_{1,...,r} = \sum_{p=1}^{r}(-1)^p(p-1)!E\left[\prod_{i\in v_1} x_i\right].E\left[\prod_{j\in v_2} x_j\right]...E\left[\prod_{k\in v_p} x_k\right],$$

où tous les ensembles $\{v_1, v_2, ..., v_p : 1 \leq p \leq r\}$ forment une partition de $\{1, 2, ..., r\}$, p est le nombre d'éléments composant la partition.

3 Quelques propriétés

a- $Cum(\alpha_1 X_1, \alpha_2 X_2, ..., \alpha_n X_n) = \alpha_1\alpha_2...\alpha_n Cum(X_1, X_2, ...X_n)$.

b- $Cum(X + Y, Z_1, ...Z_n) = Cum(X, Z_1, ..., Z_n) + Cum(Y, Z_1, ..., Z_n)$.

c- Si les variables aléatoires X_i sont indépendantes, deux à deux, on a :

$$Cum(X_1^{q_1}, X_2^{q_2}, ..., X_n^{q_n}) = 0.$$

Pour plus de détails, on se réfère à [Mccullagh87], à [Pincibono93] et à [Lacoume et al.97].

4 Estimation de statistiques d'ordre supérieur

Il y a un grand nombre d'estimateurs possibles pour une statistique donnée, le choix qui est fait ci-dessous est lié à l'utilisation que l'on veut en faire. En effet le signal étudié est en général peu stationnaire, et l'estimation de manière récursive permet de ne pas introduire de retard dans les algorithmes de détection. Les estimateurs sur des fenêtres exponentielles répondent à ces exigences. Dans toute cette section on fera l'hypothèse que les variables aléatoires étudiées sont indépendantes et indentiquement distribuées (i.i.d.). Dans le cas du signal de parole et de bruit cette hypothèse est bien sûr abusive, mais donne un ordre d'idée quant aux moyennes et variances des estimateurs de variables corrélées.

4.1 Moments d'ordre 1

L'estimation "classique" de la moyenne, est la moyenne arithmétique :

$$\hat{\mu}_1(n) = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

où x_i est une observation de la variable aléatoire X. L'estimateur $\hat{\mu}_1(n)$ est sans biais. On a $E[\hat{\mu}_1(n)] = m$ et $Var(\hat{\mu}_1(n)) = \frac{\sigma^2}{n}$, quelque soit n, où m et σ^2 sont respectivement la moyenne et la variance théoriques de la variable aléatoire X.

Pour prendre en compte le caractère non stationnaire des signaux, l'estimation de la moyenne peut se faire sur une fenêtre exponentielle, qui joue le rôle d'un filtre passe-bas avec un facteur d'oubli. Cette estimation a de plus l'avantage de se faire de façon récursive. On a ainsi :

$$\hat{\mu}_1(n) = \hat{\mu}_1(n-1) + (1-\lambda)(x_n - \hat{\mu}_1(n-1)) = (1-\lambda)\sum_{i=0}^{n} \lambda^{n-i} x_i,$$

où λ est le facteur d'oubli. On a supposé $\hat{\mu}_1(0) = 0$. La moyenne de cet estimateur est :

$$E[\hat{\mu}_1(n)] = m\,(1-\lambda^{n+1}),$$

qui est asymptotiquement sans biais. Sa variance est donnée par :

$$Var(\hat{\mu}_1(n)) \;=\; \frac{1-\lambda}{1+\lambda}\,(1-\lambda^{2(n+1)})\sigma^2,$$

qui a pour limite, lorsque $n \to +\infty$: $\frac{1-\lambda}{1+\lambda}\,\sigma^2$. Cette valeur sera d'autant plus petite que le facteur d'oubli λ sera proche de 1. Le problème est que plus le facteur d'oubli λ est proche de 1, c'est à dire plus la fenêtre exponentielle est grande, plus le choix s'applique à un signal supposé stationnaire. L'hypothèse sur le degré de stationnarité du signal détermine donc le choix du facteur d'oubli.

4.2 Statistiques d'ordre 2

De la même façon que pour l'estimation de la moyenne on peut estimer le moment d'ordre deux par :

$$\hat{\mu}_2(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2,$$

qui est un estimateur sans biais de moyenne :

$$E[\hat{\mu}_2(n)] = m^2 + \sigma^2,$$

et de variance :

$$Var(\hat{\mu}_2(n)) = \frac{1}{n}(\mu_4 - (m^2+\sigma^2)^2),$$

où μ_4 est le moment d'ordre 4 théorique de X.

L'estimation peut aussi se faire sur une fenêtre exponentielle :

$$\hat{\mu}_2(n) = \hat{\mu}_2(n-1) + (1-\lambda)(x_n^2 - \hat{\mu}_2(n-1)) = (1-\lambda)\sum_{i=0}^{n} \lambda^{n-i} x_i^2.$$

Cet estimateur a pour moyenne :

$$E[\hat{\mu}_2(n)] = (1-\lambda^{n+1})(m^2+\sigma^2),$$

qui est asymptotiquement sans biais. Sa variance est :

$$Var(\hat{\mu}_2(n)) \;=\; \frac{1-\lambda}{1+\lambda}(1-\lambda^{2(n+1)})(\mu_4 - (m^2+\sigma^2)^2).$$

On a donc :

$$\lim_{n\to+\infty} Var(\hat{\mu}_2(n)) = \frac{1-\lambda}{1+\lambda}(\mu_4 - (m^2+\sigma^2)^2).$$

Cette limite converge vers 0 lorsque le facteur d'oubli λ tend vers 1.

Remarque : On constate que asymptotiquement en n, la moyenne et la variance de l'estimateur du moment d'ordre 2 avec un facteur d'oubli λ est équivalent à la moyenne et la variance de l'estimateur arithmétique pour $n = \frac{1+\lambda}{1-\lambda}$. C'est également le cas pour le moment d'ordre 1.

L'estimation de la variance peut se faire à partir des estimateurs précédents de différentes façons. Citons :

$$\hat{\sigma}^2(n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_1(n))^2,$$

où $\hat{\mu}_1(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$. On peut montrer que :

$$E[\hat{\sigma}^2(n)] = \frac{n-1}{n}\sigma^2,$$

et,

$$Var(\hat{\sigma}^2(n)) = \frac{n-1}{n^3}[(n-1)\mu_4 - (n-3)\sigma^4] - \frac{n-1}{n^3}[(n-1)m^4 - 2(n-3)m^2\sigma^2].$$

Cet estimateur n'est pas sans biais, mais $\hat{\sigma^*}^2(n) = \frac{n}{n-1}\hat{\sigma}^2(n)$ est non biaisé. Si la valeur théorique de la moyenne est connue on peut récrire cet estimateur comme :

$$\hat{\sigma}^2(n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 - (\hat{\mu}_1(n) - m)^2.$$

$\hat{\sigma}^2(n)$ s'écrit aussi :

$$\hat{\sigma}^2(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\hat{\mu}_1(n))^2.$$

On peut de même réutiliser les statistiques sur les fenêtres exponentielles. On a ainsi :

$$\hat{\sigma}^2(n) = (1-\lambda)\sum_{i=0}^{n}\lambda^{n-i}x_i^2 - (\hat{\mu}_1(n))^2,$$

où à présent, $\hat{\mu}_1(n) = (1-\lambda)\sum_{i=0}^{n}\lambda^{n-i}x_i$. On peut obtenir la moyenne de cet estimateur :

$$\begin{aligned} E[\hat{\sigma}^2(n)] &= \left[1 - \frac{1-\lambda}{1+\lambda}(1-\lambda^{n+1})\right](1-\lambda^{n+1})\sigma^2 \\ &\quad +\lambda^{n+1}(1-\lambda^{n+1})m^2, \end{aligned}$$

qui est asymptotiquement biaisé, sa limite en n est $\frac{2\lambda}{1+\lambda}\sigma^2$. On remarque que si λ tend vers 1, cet estimateur devient non biaisé. On a de plus, la variance de cet estimateur :

$$\begin{aligned} Var(\hat{\sigma}^2(n)) = &\left[\frac{1-\lambda}{1+\lambda}(1-\lambda^{2(n+1)}) - 2\,\frac{(1-\lambda)^2}{1+\lambda+\lambda^2}(1-\lambda^{3(n+1)})\right. \\ &\left. +\frac{(1-\lambda)^3}{(1+\lambda)(1+\lambda^2)}(1-\lambda^{4(n+1)})\right]\mu_4 \\ &+\left\{4\,\frac{(1-\lambda)^2}{(1+\lambda)^2}\left[\lambda\,\frac{1-\lambda^{4(n+1)}}{1+\lambda^2} - \lambda^{2(n+1)}(1-\lambda^{2(n+1)})\right]\right. \\ &+2\,\frac{(1-\lambda)^2}{1+\lambda+\lambda^2}(1-\lambda^{3(n+1)}) - \frac{1-\lambda}{1+\lambda}(1-\lambda^{2(n+1)}) \\ &\left.-\frac{(1-\lambda)^3}{(1+\lambda)(1+\lambda^2)}(1-\lambda^{4(n+1)})\right\}(\sigma^2+m^2)^2 \\ &-4\,\frac{(1-\lambda)^2}{(1+\lambda)^2}\left[\lambda\,\frac{1-\lambda^{4(n+1)}}{1+\lambda^2} - \lambda^{2(n+1)}(1-\lambda^{2(n+1)})\right]m^4 \end{aligned}$$

On a cependant :

$$\lim_{n\to+\infty} \quad Var(\hat{\sigma}^2(n)) = \left(\frac{1-\lambda}{1+\lambda} - 2\frac{(1-\lambda)^2}{1+\lambda+\lambda^2} + \frac{(1-\lambda)^3}{(1+\lambda)(1+\lambda^2)}\right)\mu_4$$
$$+\left[4\lambda\frac{(1-\lambda)^2}{(1+\lambda)^2(1+\lambda^2)} + 2\frac{(1-\lambda)^2}{1+\lambda+\lambda^2} - \frac{1-\lambda}{1+\lambda}\right.$$
$$\left.-\frac{(1-\lambda)^3}{(1+\lambda)(1+\lambda^2)}\right](\sigma^2+m^2)^2$$
$$-4\lambda\frac{(1-\lambda)^2}{(1+\lambda)^2(1+\lambda^2)}m^4$$

Cette limite tend vers zéro lorsque λ tend vers 1. Cet estimateur est donc consitant.

En supposant que la variable aléatoire suit une loi Laplacienne, l'estimation de la variance peut se faire par l'estimation de l'écart type :

$$\begin{aligned}\hat{\sigma}_L(n) &= \hat{\sigma}_L(n-1) + (1-\lambda)(|x_n - \hat{\mu_1}(n)| - \hat{\sigma}_L(n-1)) \\ &= (1-\lambda)\sum_{i=0}^{n}\lambda^{n-i}|x_i - \hat{\mu_1}(i)|.\end{aligned}$$

L'estimation de la variance est alors :

$$\hat{\sigma}_L^2(n) = (\hat{\sigma}_L(n))^2.$$

Cet estimateur de la variance a l'avantage de ne faire appel qu'au moment d'ordre 1, il peut donc être moins coûteux en temps de calculs. C'est pourquoi l'hypothèse d'une distribution Laplacienne est parfois faite pour le calcul de la variance indépendamment des autres hypothèses.

4.3 Statistiques d'ordre supérieur à 2

De même que précédemment, on peut estimer les moments d'ordre 3 et 4, de façon arithmétique :

$$\hat{\mu}_3(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^3,$$

et

$$\hat{\mu}_4(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^4.$$

Ce sont des estimateurs sans biais. Et de la même façon, sur des fenêtres exponentielles :

$$\hat{\mu}_3(n) = (1-\lambda)\sum_{i=1}^{n} \lambda^{n-i} x_i^3,$$

et

$$\hat{\mu}_4(n) = (1-\lambda)\sum_{i=1}^{n} \lambda^{n-i} x_i^4.$$

Ces estimateurs ont pour espérance mathématique :

$$E[\hat{\mu}_3(n)] = (1-\lambda^{n+1})\mu_3,$$

et

$$E[\hat{\mu}_4(n)] = (1-\lambda^{n+1})\mu_4.$$

Ils sont ainsi asymptotiquement sans biais. Le calcul des variances des moments d'ordre 3 et 4 pose des difficultés dans l'établissement des formules générales.

On peut estimer les cumulants d'ordre 3 et 4, d'après les formules précédemment citées :

$$\hat{\kappa}_3(n) = \hat{\mu}_3(n) - 3\hat{\mu}_1(n)\hat{\mu}_2(n) + 2\hat{\mu}_1^3(n),$$

et

$$\hat{\kappa}_4(n) = \hat{\mu}_4(n) - 4\hat{\mu}_3(n)\hat{\mu}_1(n) - 3\hat{\mu}_2^2(n) + 6\hat{\mu}_2(n)\hat{\mu}_1^2(n) - \hat{\mu}_1^4(n).$$

Dans le cas d'une variable aléatoire centrée, pour les estimateurs arithmétiques, on a :

$$E[\hat{\kappa}_4(n)] = \kappa_4 - \frac{3}{n}(\kappa_4 + 2\mu_2^2).$$

Cet estimateur est donc asymptotiquement sans biais, et a pour variance :

$$\begin{aligned} Var(\hat{\kappa}_4(n)) &= \frac{1}{n}(\kappa_8 + 16\kappa_6\kappa_2 + 48\kappa_5\kappa_3 + 34\kappa_4^2 \\ &\quad +72\kappa_4\kappa_2^2 + 144\kappa_3^2\kappa_2 + 24\kappa_2^4). \end{aligned}$$

La variance de cet estimateur converge vers zéro lorsque n tend vers l'infini, il est donc consistant.

De même le skewness et le kurtosis peuvent être estimés par :

$$\hat{\chi}(n) = \frac{\hat{\kappa}_3(n)}{(\hat{\sigma}^2(n))^{\frac{3}{2}}},$$

et

$$\hat{\gamma}(n) = \frac{\hat{\kappa}_4(n)}{(\hat{\sigma}^2(n))^2}.$$

Ces estimateurs ont été étudiés, seulement de façon approchée, pour les grandes valeurs de n, par exemple dans [Mccullagh87]. Il est montré qu'ils sont biaisés au premier ordre (leur biais dépend des cumulants d'ordre plus élevé), et qu'ils sont corrélés. Il existe cependant des résultats exacts, dans le cas où la variable aléatoire est centrée et suit une loi gaussienne, on a :

$$\begin{aligned} E[\hat{\chi}(n)] &= 0, \\ E[\hat{\gamma}(n)] &= 0, \\ Var(\hat{\chi}(n)) &= \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \simeq \frac{6}{n}, \\ Var(\hat{\gamma}(n)) &= \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)} \simeq \frac{24}{n}. \end{aligned}$$

Il apparaît que plus le moment (et donc le cumulant) est d'ordre important, plus la variance de son estimateur sera grande.

Une autre quantité qui peut être intéressante, est le moment, non centré, mais normalisé par la variance, i.e. :

$$\hat{m}_3(n) = \frac{\hat{\mu}_3(n)}{(\hat{\sigma}^2(n))^{\frac{3}{2}}},$$

et

$$\hat{m}_4(n) = \frac{\hat{\mu}_4(n)}{(\hat{\sigma}^2(n))^2}.$$

Ces estimateurs ont une variance plus faible que le skewness et le kurtosis. Bien sûr, si la variable aléatoire considérée est centrée, on a :

$$\hat{\chi}(n) = \hat{m}_3(n),$$

et

$$\hat{\gamma}(n) = \hat{m}_4(n) - 3.$$

Les résultats théoriques sont souvent simplifiés, si les variables aléatoires sont centrées, or ceci n'est pas toujours le cas. Pour effectuer un centrage de manière parfaite, il faut considérer un grand nombre de données.

Sous l'hypothèse de variables i.i.d., les moments et cumulants d'ordre supérieur estimés ci-dessus, sont donc biaisés, asymptotiquement biaisés, mais lorsque λ tends vers 1, les estimateurs ne sont plus biaisés. De plus leur variance converge vers zéro lorsque le nombre d'échantillons augmente, ils sont

Moyenne: $n = 1000$, test sur 998 échantillons					
	$\lambda = 0.9$	$\lambda = 0.99$	$\lambda = 0.995$	$\lambda = 0.999$	théorique
$\hat{\mu}_1(n)$	0.5006	0.5004	0.5006	0.5004	0.5004
$\hat{\mu}_2(n)$	0.3330	0.3334	0.3337	0.3337	0.3337
$\hat{\sigma}^2(n)$	0.0781	0.0826	0.0830	0.0833	0.0833
$\hat{\sigma}_L^2(n)$	0.0590	0.0613	0.0641	0.2433	0.0833
$\hat{\chi}(n)$	−0.0061	−0.0012	−0.0015	−0.0001	0.0000
$\hat{m}_3(n)$	12.1429	10.5749	10.4952	10.4236	10.4152

TAB. 1 - *Moyenne expérimentale pour* $n = 1000$

donc consistants. Ces estimateurs peuvent donc nous permettre de mieux caractériser les distributions des variables étudiées.

Les moyennes et les variances de ces estimateurs étant difficiles à établir théoriquement, on va, à présent, étudier expérimentalement les moyennes et les variances de quelques estimateurs.

5 Estimation numérique de la moyenne et de la variance de quelques estimateurs

Le calcul théorique de la moyenne et de la variance des estimateurs des statistiques d'ordre supérieur devient vite complexe. C'est pourquoi, on présente dans les tableaux 1, 3 et 4, quelques résultats expérimentaux, sur les estimateurs calculés à partir de fenêtres exponentielles, vus précédemment. Ces estimateurs ont été initialisés par leur valeur estimée arithmétiquement. Les calculs ont été faits à partir d'une variable aléatoire de loi uniforme $[0, 1]$, générée avec le générateur de Matlab. Dans un premier temps les calculs ont été faits pour 998 échantillons, avec une valeur de $n = 1000$. Dans la colonne "théorique" a été calculé les valeurs des estimateurs, par une estimation arithmétique.

On remarque que plus l'ordre de la statistique estimée est important, plus il est nécessaire de prendre un facteur d'oubli proche de 1, pour une même précision de la moyenne.

Dans le but de leur comparaison, les valeurs numériques de la variance, sont calculées, comme une proportion de la moyenne. C'est-à-dire, la variance observée a été divisée par la moyenne observée, puis multipliée par 100. Dans le tableau 2 figurent les valeurs théoriques de cette proportion calculée à partir des formules théoriques présentées dans le paragraphe précédent. Les

Variance théorique				
	$\lambda = 0.9$	$\lambda = 0.99$	$\lambda = 0.995$	$\lambda = 0.999$
$\hat{\mu}_1(n)$	0.88	0.08	0.04	0.01
$\hat{\mu}_2(n)$	1.41	0.13	0.07	0.01

TAB. 2 - *Variance à partir des formules théoriques*

Variance : $n = 1000$, test sur 998 échantillons				
	$\lambda = 0.9$	$\lambda = 0.99$	$\lambda = 0.995$	$\lambda = 0.999$
$\hat{\mu}_1(n)$	0.86	0.08	0.04	0.01
$\hat{\mu}_2(n)$	1.38	0.13	0.06	0.01
$\hat{\sigma}^2(n)$	0.39	0.03	0.02	0.0028
$\hat{\sigma}^2_L(n)$	0.37	0.04	0.02	0.036
$\hat{\chi}(n)$	1962.30	858.33	346.67	844.21
$\hat{m}_3(n)$	179.80	9.72	4.82	0.86

TAB. 3 - *Variance expérimentale pour* $n = 1000$

valeurs "théoriques" de la moyenne, variance, et autres moments, nécessaires pour l'obtention de ces résultats ont été calculées à partir des estimateurs arithémétiques.

La valeur pour $\lambda = 0.999$ peut être erratique, ceci est dû au fait que la taille de la fenêtre correspond à n. En effet, pour $\lambda = 0.999$, l'estimation se fait sur $N = \frac{1}{1-\lambda} = 1000$ valeurs. Pour remédier à ce problème, et pour calculer la variance avec des facteurs d'oubli plus importants, on présente les tableaux 2, 3 et 4, le calcul fait avec $n = 10000$ et un nombre moins important d'échantillons. On remarque que plus l'ordre de la statistique estimée est faible, moins il est nécessaire d'avoir un facteur d'oubli proche de 1 pour obtenir un taux de variance faible. De plus pour un même facteur d'oubli, $\hat{m}_3(n)$ a une variance plus faible que $\hat{\chi}(n)$, alors qu'ils sont du même ordre.

Variance : $n = 10000$, test sur 98 échantillons			
	$\lambda = 0.999$	$\lambda = 0.9995$	$\lambda = 0.9999$
$\hat{\chi}(n)$	42.86	46.46	23.17
$\hat{m}_3(n)$	1.07	0.57	0.09

TAB. 4 - *Variance expérimentale pour* $n = 10000$

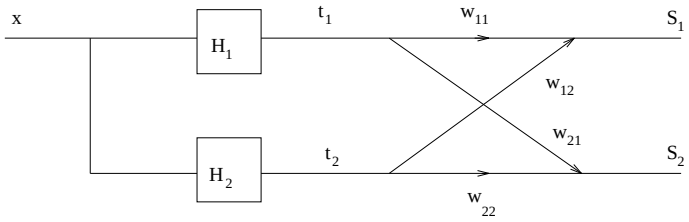


FIG. 1 - *Séparation de source*

Nous allons à présent voir comment ces statistiques peuvent être utilisées en vue du problème de détection de la parole.

6 Utilisation des statistiques d'ordre supérieur pour la détection de la parole

[Jacovitti et al.91] proposent l'utilisation du skewness et du kurtosis dans une perspective de segmentation en sons voisés-non voisés du signal. Ils supposent le signal stationnaire et estiment $\chi(n)$ et $\gamma(n)$ du signal sur des fenêtres rectangulaires. Les moments d'ordre supérieur étant instables, ils considèrent des échantillons de l'ordre de 50 msec, pour l'estimation. Ceci entraîne un retard dans la détection.

[Doukas et al.97] utilisent le fait que le cumulant croisé de deux variables aléatoires indépendantes est nul, pour discriminer le signal de parole et celui du bruit à la source. Pour obtenir ce cumulant croisé, il est nécessaire d'avoire deux sources. Ils filtrent donc la source unique à l'aide d'un filtre passe-bas H_1 et d'un filtre passe-haut H_2 (cf figure 1), et introduisent ainsi une deuxième source fictive à l'aide d'un réseau de neurones. Les filtres utilisés sont :

$$\begin{aligned} H_1 &= \frac{1}{2} + z^{-1} + \frac{1}{2}z^{-2}, \\ H_2 &= -\frac{1}{2} + z^{-1} - \frac{1}{2}z^{-2}. \end{aligned}$$

Ils cherchent ensuite à minimiser la fonction :

$$J = \sum_{i,j=0}^{+\infty} \left(E[s_1^{2i+1} s_2^{2j+1}]\right)^2.$$

En pratique la somme s'effectue pour un grand nombre de i et j. Lorsque le signal est stationnaire J reste constante, et lorsque les statistiques du signal

x changent J devient très grande. J est estimée récursivement sur une fenêtre exponentielle :

$$\hat{J}(n) = \hat{J}(n-1) + (1-\lambda)\left(\sum_{i,j=0}^{+\infty}\left(E[s_1^{2i+1}s_2^{2j+1}]\right)^2 - \hat{J}(n-1)\right).$$

Cet estimateur est comparé à un seuil T_n qui est adapté par :

$$T_n = T_{n-1} + (1-\lambda_i)(\hat{J}(n-1) - T_{n-1}),$$

où λ_i pour $i = 0, 1$, est un facteur d'oubli différent selon l'état de silence ($i = 0$) ou de parole ($i = 1$) dans lequel on est. λ_0 et λ_1 sont choisis de manière à avoir $\frac{\lambda_0}{\lambda_1} = 100$. Le changement d'état se fera lorsque $\hat{J}(n) > 1 * T_n$, lorsque l'on est dans l'état de parole, et lorsque $\hat{J}(n) < 1.4 * T_n$, lorsque l'on est dans l'état de silence.

F. Bouteille (cf [Bouteille99]), a repris cette méthode de détection d'activité vocale, simplifiant les calculs, conservant une bonne détection. Les valeurs de t_1 et de t_2 sont reprises directement sans faire appel au réseau de neurones. Pour éviter les calculs d'espérance mathématique, t_1 et t_2 sont intégrés dans $\hat{J}(n)$ de la façon suivante :

$$\hat{J}(n) = \hat{J}(n-1) + (1-\lambda)\left(\sum_{i,j} t_1^i t_2^j - \hat{J}(n-1)\right),$$

où $(i,j) \in \{(u,v) \in \mathbb{R}^2 : 0 \leq u \leq r, 0 \leq v \leq r \text{ et } u+v = r\}$, r étant l'ordre maximum des moments considérés.

Nous allons à présent voir comment intégrer les statistiques d'ordre supérieur à l'algorithme de détection Bruit/Parole utilisé au CNET.

7 Moment d'ordre 3 appliqué à l'algorithme de DBP du CNET

Rappelons que l'algorithme de DBP du CNET utilise un automate à cinq états, décrit dans [Mauuary94] (cf Annexe A), qui utilise les statistiques de l'énergie du second ordre pour décider d'un changement d'état (cf [Karray98b]). L'hypothèse selon laquelle le bruit suit une loi normale de paramètres (μ, σ^2), est faite. Les statistiques du bruit sont estimées à chaque passage dans l'état *silence* de l'automate. La moyenne est estimée par :

$$\hat{\mu}_1(n+1) = \hat{\mu}_1(n) + (1-\lambda)(energie - \hat{\mu}_1(n)),$$

et l'écart-type par :

$$\hat{\sigma}(n+1) = \hat{\sigma}(n) + (1-\lambda)(|energie - \hat{\mu}_1(n)| - \hat{\sigma}(n)),$$

où n est le numéro de l'échantillon dans l'état *silence* de l'automate. L'énergie de chaque trame est considérée, et on cherche à vérifier l'hypothèse que l'on est dans l'état *silence*, qui correspond au bruit seul. La décision sera prise en fonction de l'écart de l'énergie de cette trame par rapport à la moyenne estimée de l'énergie du bruit, c'est-à-dire selon la valeur du rapport critique $r(x) = \frac{x-\hat{\mu}_1}{\hat{\sigma}}$, comparé à un seuil. En prenant un intervalle de confiance de 95 %, ce rapport est comparé à 1.7 (cf [Karray98b] et [Karray98a]).

Nous avons choisi d'intégrer $\hat{m}_3(n)$, le moment d'ordre 3 non centré, mais normalisé par la variance. En effet on a vu que la variance de cet estimateur est plus faible que celle du skewness, de plus $\hat{m}_4(n)$, le moment normalisé d'ordre 4, non centré, a une variance encore plus importante et ne semble pas apporter beaucoup plus. La figure 2 montre les fonctions de répartition du rapport des moments d'ordre 3 et 4 calculés dans les périodes de parole sur ceux calculés dans les périodes de bruit. Les moments, non centrés, normalisés ont été calculés sur une fenêtre exponentielle. La figure 2 représente le nombre de fichiers dont le rapport des moments d'ordre 3 et 4 est inférieur à une borne supérieure donnée (variant de 0 à 3). Cette figure indique que pour un même facteur d'oubli $\lambda = 0.995$, le rapport du moment d'ordre 4 n'apporte pas une nette amélioration de la capacité de discrimination du bruit et de la parole.

Le moment, non centré, normalisé, d'ordre 3 a été calculé pour les premiers coefficients cepstraux, dans les périodes de silence et de parole, selon tous les environnements sur la base de donnée du GSM. Seul pour l'énergie, il apparaît discriminant. Nous avons donc tenté d'affiner la discrimination existante de l'énergie du bruit et de la parole. Le critère utilisé est le rapport de $\hat{m}_3(n)$ à court-terme (calculé avec un facteur d'oubli faible, $\lambda = 0.9$) et de $\hat{m}_3(n)$ à long-terme(calculé avec un facteur d'oubli plus important $\lambda = 0.995$). Le moment à long-terme est estimé dans les états *silence* de l'automate. Dans un premier temps le rapport est comparé à un seuil fixe dans les périodes détectées comme étant de la parole, avec l'algorithme initial (dbp-v2), utilisant les statistiques du second ordre. C'est-à-dire l'algorithme détectera une période de parole si le seuil de l'algorithme initial est supérieur à l'énergie de la trame courante et si le rapport des moments est inférieur au seuil fixe. Sinon l'algorithme détectera du silence.

Dans un second temps, au lieu du seuil fixe, le rapport est comparé à un seuil adaptatif, de la même façon, dans les périodes de parole que l'algorithme

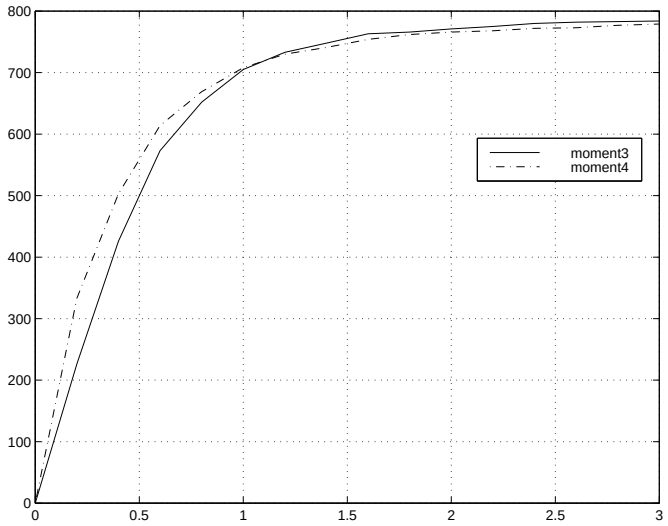


FIG. 2 - *Rapport Parole/Bruit des moments, non centrés, normalisés, d'ordre 3 et 4*

initial a détectées. Le seuil adaptatif est calculé à partir de la moyenne du rapport, multiplié par un coefficient, dans les périodes de parole, prise sur une fenêtre exponentielle. Le fait de multiplier le rapport permet d'obtenir une borne supérieure, pour le rapport des moments d'ordre 3, $\hat{m}_3(n)$, à court-terme et à long-terme. C'est-à-dire le seuil adaptatif est calculé par la formule de récurrence suivante :

$$\hat{T}(n+1) = \hat{T}(n) + (1 - \lambda_T)(coef.rap(n) - \hat{T}(n)),$$

où $rap(n)$ est le rapport des moments d'ordre 3 à court-terme et à long-terme, à l'instant n dans une période de parole, $\lambda_T = 0.99$ est le facteur d'oubli, et $coef$ est le coefficient permettant d'obtenir la borne supérieure du rapport des moments d'ordre 3 dans les périodes de parole. L'algorithme est décrit en annexe B.

Il est clair que plus le facteur d'oubli du moment à court-terme est important, plus on observera un retard sur la détection des segments de parole. Le problème réside dans le fait que la variance de cet estimateur reste grande. Il s'avère en fait que seul le moment à court-terme est discriminant.

Dans ce qui suit, on compare les trois algorithmes, l'algorithme initial, l'algorithme intégrant le rapport des moments avec le seuil fixe, et celui avec le seuil adaptatif.

8 Expérimentations

Les performances des trois algorithmes ont été comparées de différentes manières. Dans un premier temps, on étudie les pourcentages de mots bien placés lors de la segmentation, puis sur quelques fichiers, on regardera où se situe la segmentation des différents algorithmes en comparaison avec la valeur du rapport des moments d'ordre 3. Les tests sur la segmentation représentant les différents types d'erreurs, puis les tests de reconnaissance seront alors abordés.

Dans les trois tableaux 5, 6 et 7, sont représentés les pourcentages de mots détectés comme étant de la parole, qui sont bien placés, (c'est-à-dire bien segmentés), tronqués à gauche, à droite, ou à gauche et à droite, ainsi que le total des erreurs de parole, qui représente les erreurs d'omission, d'insertion, de regroupement de parole et de non-parole et les fragmentations. La dernière colonne représente le total des erreurs sur le bruit. Chaque tableau donne les valeurs pour un algorithme : 1- l'algorithme initial, utilisant les statistiques du second ordre, 2- l'algorithme avec le rapport des moments comparé à un seuil fixe, et 3- celui utilisant le seuil adaptatif. Pour ces deux derniers algorithmes, le seuil de prédétection de l'algorithme initial est fixé à 1.7. Les pourcentages ont été calculés sur tous les fichiers de la base de données GSM (cf annexe C), ils correspondent au nombre de segments de parole détectés bien placés, tronqués à gauche, à droite, ou à gauche et à droite, par rapport au nombre de segments total de parole détectés comme étant de la parole. Les pourcentages du total des erreurs de parole ont été calculés en faisant la somme des erreurs d'omission, d'insertion, des regroupements de parole et de non-parole (appartés), et des fragmentations, pour les segments de parole et de non-parole, le tout étant divisé par la somme des segments de référence (détection manuelle) de parole et de non-parole. De la même façon ont été calculés les pourcentages du total des erreurs de bruit. Ces tableaux montrent que le choix du seuil ne doit pas être fait uniquement en fonction du pourcentage de mots bien placés. Il faut touver le seuil donnant à la fois un fort nombre de mots bien placés et peu d'erreurs sur la parole et sur le bruit.

Les figures 3 et 4 représentent le rapport des moments d'ordre 3, non centrés, normalisés, calculés par l'algorithme avec le seuil fixé à 300000, et par l'algorithme avec un seuil adaptatif, le coefficient étant fixé à 3. Les rapports sont représentés pour deux fichiers, l'un contenant un bruit de fond l'autre de l'écho. Sur ces courbes sont représentées les segmentations manuelles, avec celles obtenues par l'algorithme initial, dbp-v2, et avec celles obtenues par l'algorithme avec le seuil fixe, sur la figure 3, et avec le seuil adaptatif, sur la figure 4. On remarque que dans les deux cas les courbes sont sensiblement

Résultats (en %) avec l'algorithme initial						
seuil	bien placé	tr. à g.	tr. à d.	tr. à g. et à d.	err. P	err. B
1.9	71.17	11.94	9.43	5.34	15.32	24.46
1.8	73.61	10.93	8.68	4.64	15.38	27.74
1.7	76.02	9.93	7.86	3.96	15.60	30.64
1.6	78.40	9.12	6.80	3.27	16.11	33.65
1.5	80.58	8.29	5.81	2.62	17.01	36.66

TAB. 5 - *Résultats de la segmentation en %, avec l'algorithme initial*

Résultats (en %) avec le seuil fixe						
seuil	bien placé	tr. à g.	tr. à d.	tr. à g. et à d.	err. P	err. B
250000	66.25	11.47	14.68	6.40	16.57	24.81
300000	68.51	11.47	12.76	5.88	16.06	25.71
350000	70.26	11.21	11.57	5.51	15.7	26.46
400000	71.40	11.11	10.80	5.25	15.49	27.27
1000000	75.12	10.36	8.41	4.19	15.36	29.49

TAB. 6 - *Résultats de la segmentation en %, avec le seuil fixe*

Résultats (en %) avec le seuil adaptatif						
coefficient	bien placé	tr. à g.	tr. à d.	tr. à g. et à d.	err. P	err. B
2	60.72	12.22	19.77	6.06	15.22	23.79
3	70.33	11.36	11.88	4.87	15.17	27.12
4	73.54	10.88	9.39	4.48	15.27	28.49
5	74.52	10.7	8.64	4.27	15.23	28.97
6	75.14	10.46	8.28	4.15	15.23	29.31

TAB. 7 - *Résultats de la segmentation en %, avec le seuil adaptatif*

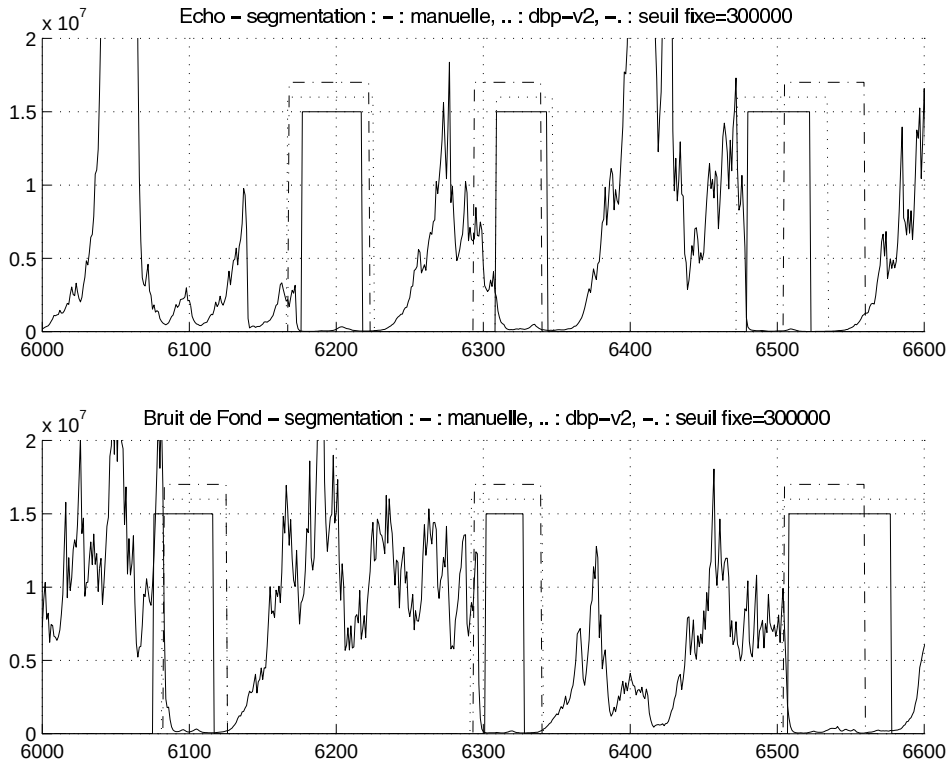


FIG. 3 - *Évaluation de la segmentation sur 2 fichiers, l'un contenant de l'écho, l'autre du bruit de fond, avec l'algorithme à seuil fixe.*

identiques, la différence entre les deux algorithmes n'est pas flagrante. La segmentation par ces algorithmes entraîne parfois une coupure en fin ou début de mots. L'algorithme utilisant le seuil adaptatif (figure 4) semble légèrement meilleur.

Sur la figure 5, les résultats des tests de segmentation présentés, sont obtenus par comparaison avec la segmentation manuelle. Parmi les erreurs comptabilisées, certaines peuvent être corrigées par une procédure de rejet dans le système de reconnaissance (par exemple les détections de bruit ou d'apparté), et d'autres ne peuvent pas être corrigées (par exemple les omissions de parole). On représente ainsi les erreurs *définitives*, composées des omissions, des regroupements et des fragmentations de parole, en fonction des

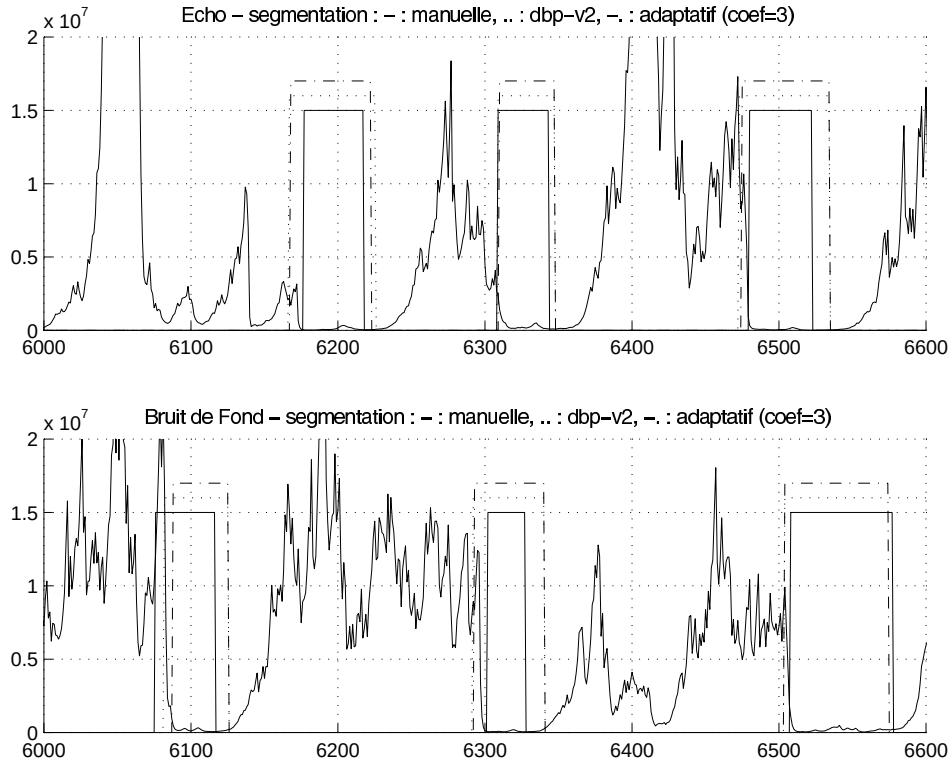


FIG. 4 - *Évaluation de la segmentation sur 2 fichiers, l'un contenant de l'écho, l'autre du bruit de fond, avec l'algorithme à seuil adaptatif.*

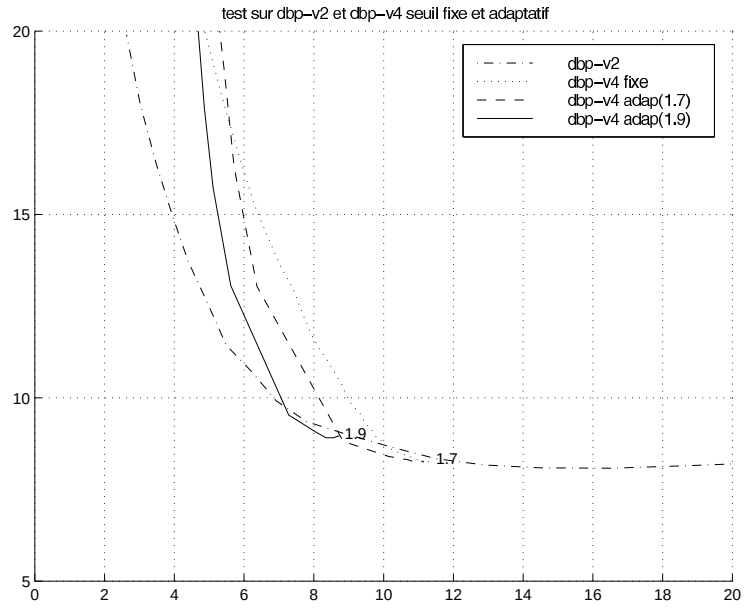


FIG. 5 - *Évaluation de l'algorithme de détection Bruit/Parole avec l'introduction du rapport des moments d'ordre 3.*

erreurs *rejetables*, composées des insertions et des détections de non parole (les apartés). Dans le cas d'une détection correcte, la détection est contabilisée "bien placée", "tronquée à gauche", ou "tronquée à droite ou tronquée à gauche et à droite". Faire varier un paramètre d'un algorithme, revient, par exemple, à augmenter les erreurs définitives et diminuer les erreurs rejetables. Les courbes ainsi tracées permettent la comparaison de plusieurs algorithmes. Pour d'avantage de précisions sur cette méthode d'évaluation, on se réfère à [Mauuary94].

La figure 5 représente les résultats de ces tests de segmentation, pour l'algorithme initial, l'algorithme avec le rapport comparé à un seuil fixe (que l'on fait varier), le seuil de l'algorithme initial étant fixé à 1.7, et l'algorithme avec le rapport comparé à un seuil adaptatif (on fait varier le coefficient dans le calcul du seuil adaptatif), le seuil de l'algorithme initial étant fixé à 1.7 et à 1.9. Les tests ont été effectués sur la base de données des enregistrements GSM, sur tous les environnements. Sur certains points, on remarque une légère amélioration, avec le seuil fixe et le seuil adaptatif. Cependant cette amélioration tend à s'annuler avec l'acroissement du seuil, les résultats se rapprochant, naturellement de l'algorithme initial.

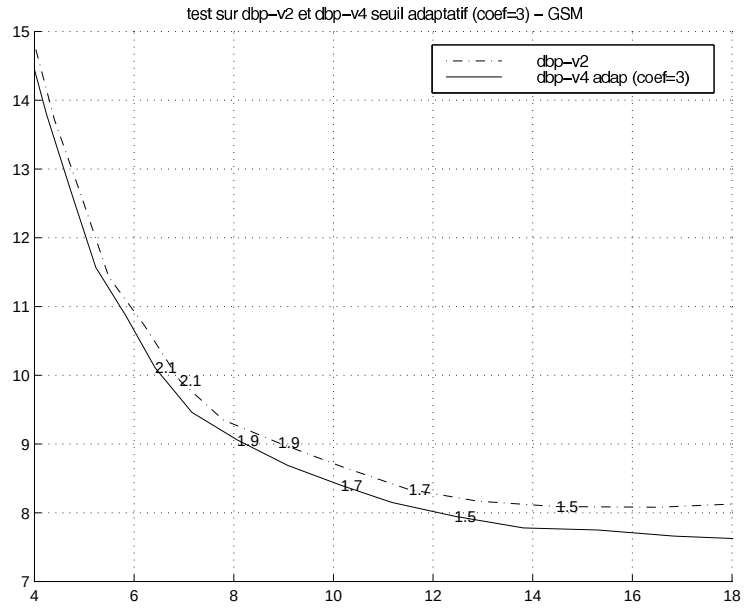


FIG. 6 - *Évaluation de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement GSM*

La figure 6 représente les résultats des tests de segmentation, pour l'algorithme initial, et pour l'algorithme avec le rapport comparé à un seuil adaptatif, où le coefficient multiplicatif donnant la borne supérieure du rapport est fixé à 3. On fait varier ici, le seuil de l'algorithme initial. On remarque que pour un seuil donné, le nombre d'erreurs définitives est légèrement inférieur à celui obtenu par l'algorithme initial, et le nombre d'erreurs rejetables est aussi inférieur.

La figure 7, représente les mêmes résultats, mais dans un environnement RTC (cf annexe C). On constate qu'au contraire des résultats précédents, l'algorithme initial est légèrement meilleur que celui utilisant les moments d'ordre 3 avec un seuil adaptatif. Dans ce cas on remarque que pour un seuil donné, le nombre d'erreurs définitives est supérieur, mais le nombre d'erreurs rejetables est légèrement inférieur.

Dans les figures 8 et 9 sont représentés les résultats des tests de reconnaissance. Dans la figure 8, on a représenté les erreurs de substitution en

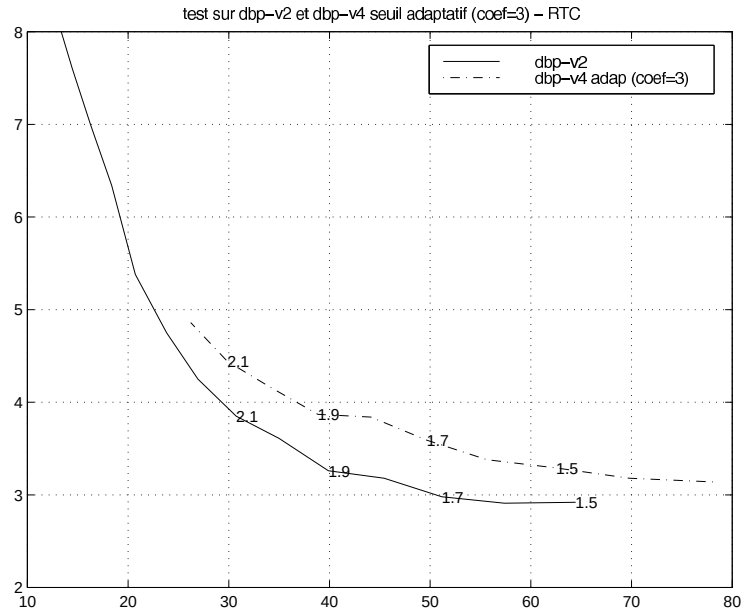


FIG. 7 - *Évaluation de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement RTC*

fonction des erreurs de faux rejets, tandis que dans la figure 9, ce sont les erreurs de fausse acceptation en fonction des erreurs de faux rejets qui sont représentés. Les tests ont été effectués à partir de deux seuils 1.5 et 1.7 pour l'algorithme initial dbp-v2 et l'algorithme utilisant le critère des moments d'ordre 3, avec un seuil adaptatif et un coefficient fixé à 3. On remarque que les résultats pour les deux algorithmes sont proches, avec néanmoins de meilleurs résultats pour l'algorithme initial.

En annexe E, on présente les résultats des tests de reconnaissance par environnement.

Cependant, les figures 10 et 11 représentant les résultats des tests de reconnaissance avec un modèle flexible (cf annexe D), montrent une légère amélioration des résultats avec l'algorithme utilisant le rapport des moments d'ordre 3, qu'avec l'algorithme initial. Cette amélioration est sans doute liée au fait que l'algorithme utilisant les moments d'ordre 3 donne moins d'erreurs définitives.

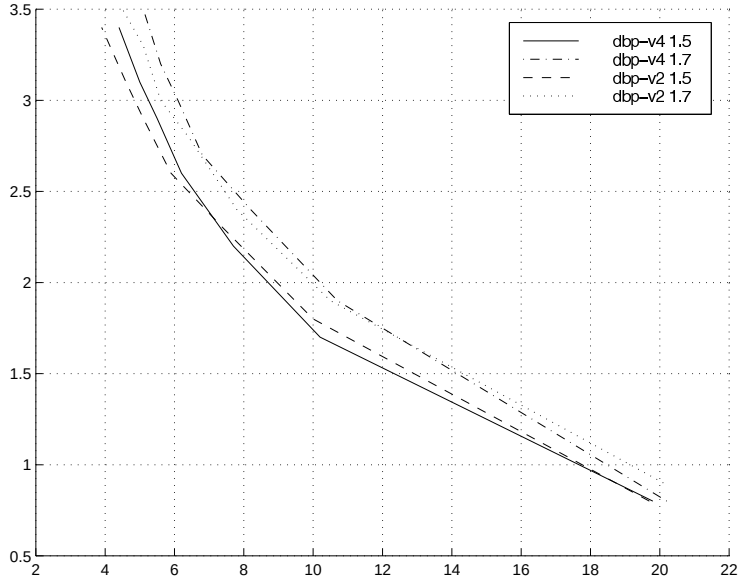


FIG. 8 - *Évaluation par les tests de reconnaissance de l'agorithme adaptatif avec un coefficient fixé à 3, des erreurs de substitution.*

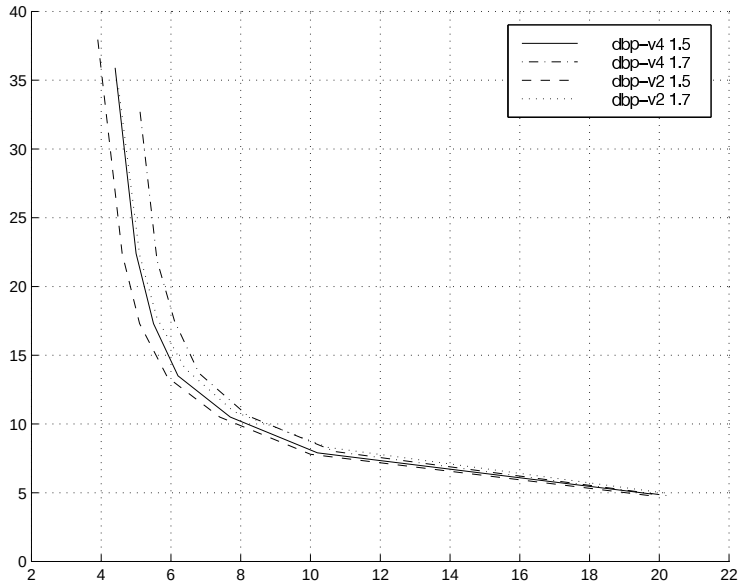


FIG. 9 - *Évaluation par les tests de reconnaissance de l'agorithme adaptatif avec un coefficient fixé à 3, des fausses alarmes.*

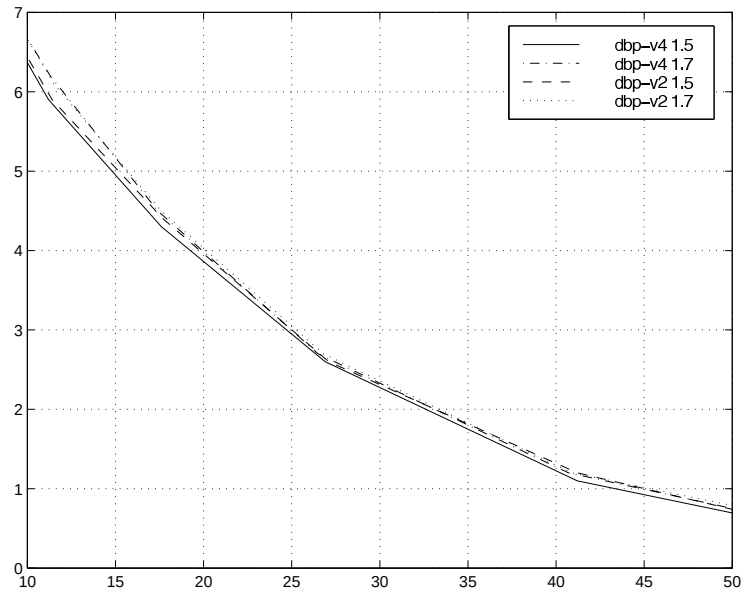


FIG. 10 - *Évaluation par les tests de reconnaissance avec un modèle flexible de l'agorithme adaptatif avec un coefficient fixé à* 3*, des erreurs de substitution.*

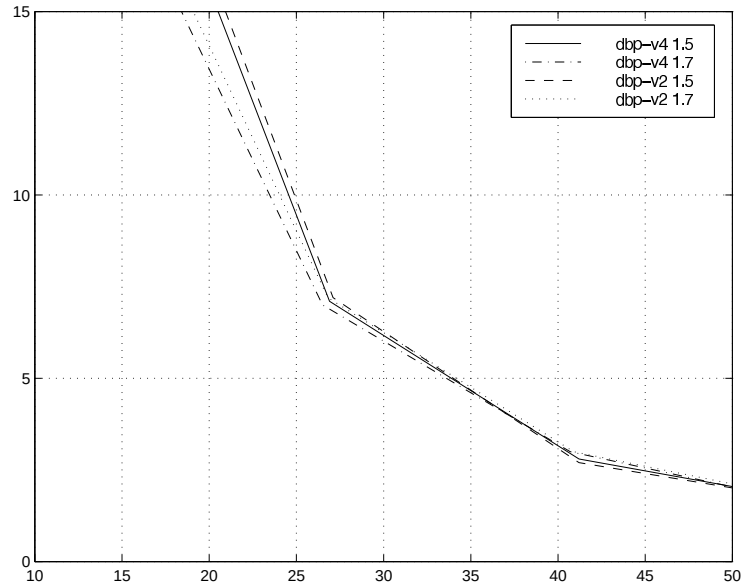


FIG. 11 - *Évaluation par les tests de reconnaissance avec modèle flexible de l'agorithme adaptatif avec un coefficient fixé à* 3*, en fonction des fausses alarmes.*

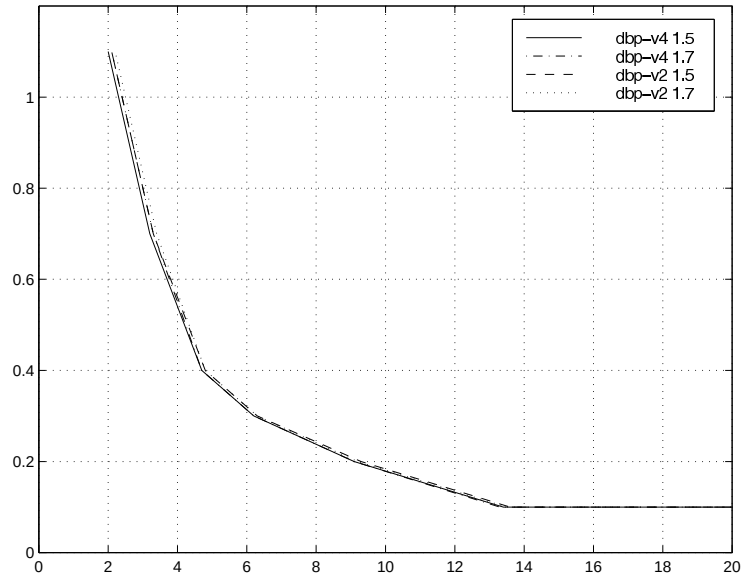


FIG. 12 - *Évaluation par les tests de reconnaissance de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement RTC.*

On présente de même dans les figures 12, 13, 14 et 15, les résultats des tests de reconnaissance avec un modèle flexible, et sans, sous environnement RTC (cf annexe C). On remarque que les résultats sont ici aussi sensiblement identiques, avec cependant une légère amélioration avec un modèle flexible.

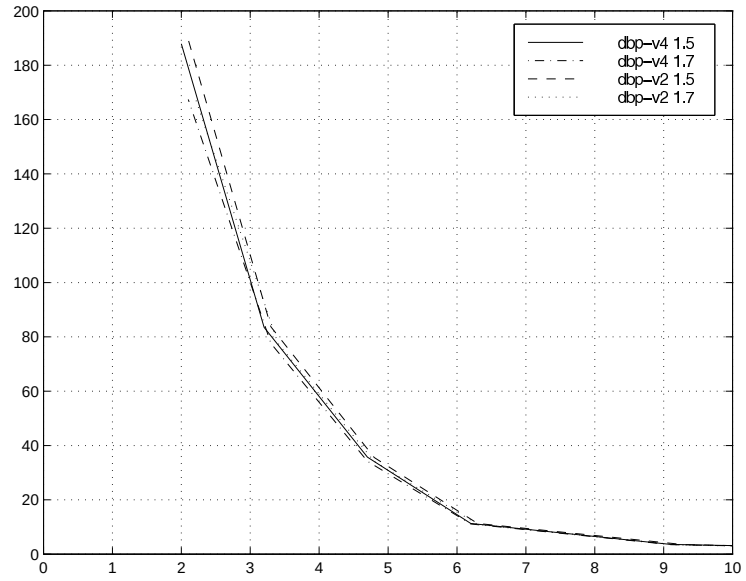


FIG. 13 - *Évaluation par les tests de reconnaissance de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement RTC.*

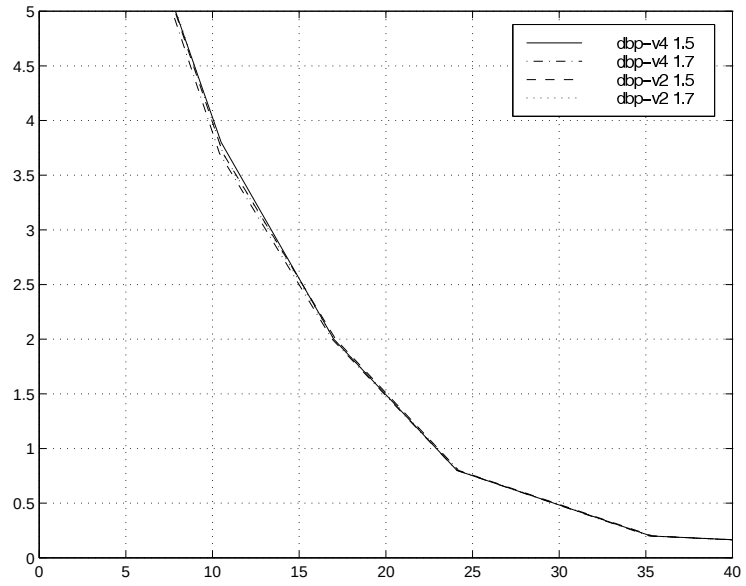


FIG. 14 - *Évaluation par les tests de reconnaissance avec un modèle flexible de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement RTC.*

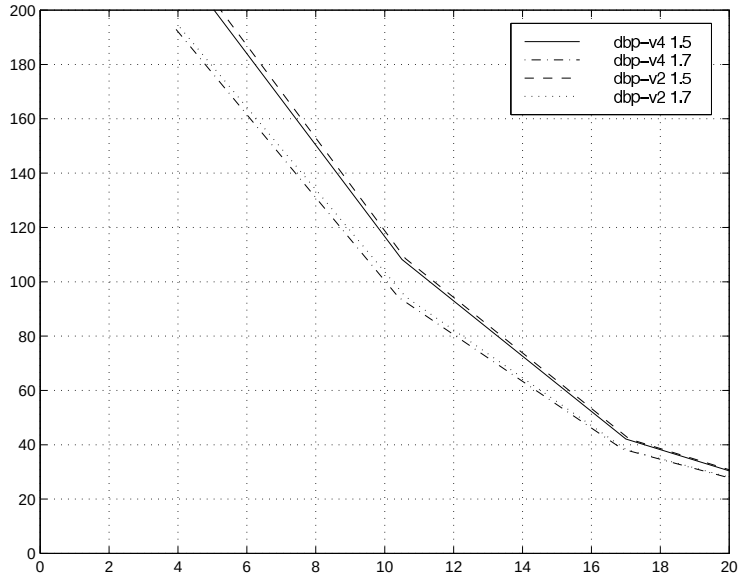


FIG. 15 - *Évaluation par les tests de reconnaissance avec un modèle flexible de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement RTC.*

9 Conclusion

Les expérimentations faites montrent que l'énergie paraît le coefficient le plus discriminant, et l'algorithme actuel semble bien l'exploiter. L'introduction des moments d'ordre supérieur n'apporte qu'une trop faible amélioration pour nous permettre de retenir ce critère. En effet, le coût de calcul et le retard, même faible, qu'entraîne l'estimation du rapport des moments, ne justifient pas son introduction, au vue des résultats obtenus. Le moment, non centré, normalisé, d'ordre 3 reste difficile à estimer sur de courtes périodes, sa variance étant trop importante. Calculer ce moment directement à partir du signal pourrait être une alternative à ce problème. Il deviendrait ainsi un critère représentant plus les caractéristiques du signal.

ANNEXES

A Description du fonctionnement de l'automate initial

Le système de détection Bruit/Parole du CNET utilise un automate à cinq états, qui sont :

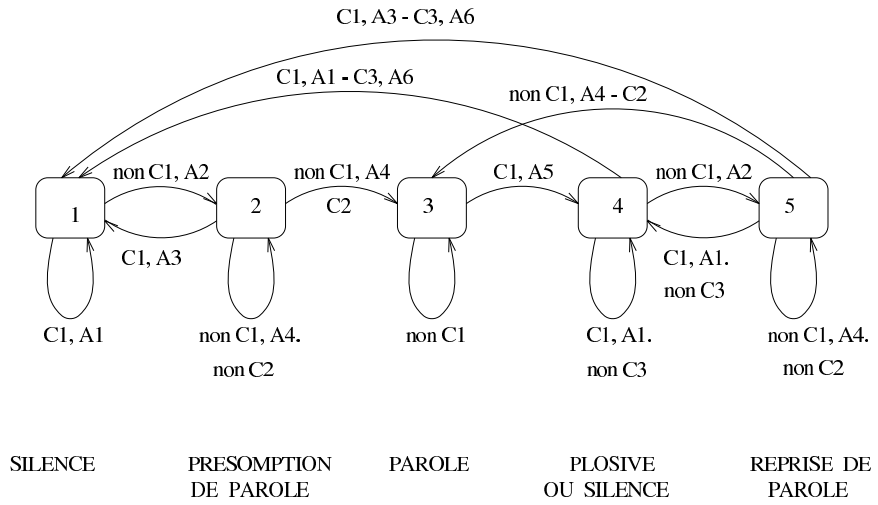
silence, présomption de parole, parole, plosive ou silence, et reprise possible de parole.

Dans une première version les passages d'un état à l'autre sont conditionnés par un seuil sur l'énergie du signal et par des contraintes structurelles de durée (durée minimum d'une voyelle et durée maximum d'une plosive). D'autres méthodes de passage d'un état à l'autre ont été étudiées, et sont présentées ci-dessous. Les passages à l'état *parole* déterminent les frontières de la parole dans le signal. Le système de reconnaissance prend en compte ces données avec une marge de sécurité sur les frontières.

L'état *silence* est l'état initial de l'algorithme. On fait ainsi l'hypothèse que la communication débute par du silence. Le détecteur reste dans cet état tant qu'il n'y a pas de trame énergétique (i.e. une trame dont l'énergie est supérieure au seuil). À la première trame énergétique, le détecteur passe dans l'état *présomption de parole*. Dans cet état, une trame non énergétique le fait retourner à l'état *silence*. Après être resté un nombre de trames minimum dans l'état *présomption de parole*, le détecteur passe à l'état *parole*. Il y reste tant que les trames sont énergétiques. Il passera à l'état *plosive ou silence*, dès que la trame sera non énergétique. Il faut au moins un certain nombre de trames non énergétiques pour confirmer le silence et retourner dans l'état *silence*, sinon le détecteur passe dans l'état *reprise possible de parole*. Dans cet état, une trame non énergétique le fait retourner dans l'état *plosive ou silence* ou dans l'état *silence* si la durée cumulée du temps passé dans l'état *plosive ou silence* et dans l'état *reprise possible de parole* représente au moins un certain nombre de trames. Après être resté un nombre de trames énergétiques minimum dans l'état *reprise possible de parole*, le détecteur retourne dans l'état *parole*.

Cet algorithme est décrit en Fig. 16.

Pour plus de détails sur le fonctionnement de l'automate, on se réfère à [Mauuary94].



CONDITIONS :

C1 : Energie < Seuil

C2 : Durée Parole (DP) >= Parole Minimum

C3 : Durée Silence (DS) >= Silence Fin

ACTIONS :

A1 : DS = DS +1

A2 : DP = 1

A3 : DS = DS + DP

A4 : DP = DP + 1

A5 : DS = 1

A6 : DS = DP = 0

VALEURS INITIALES :

ETAT = SILENCE

DS = DP = 0

Silence Fin est généralement choisi entre 240 ms (mots isolés) et 640 ms (mots connectés),
Silence Fin représente le maximum entre la durée maximum d'une tenue de plosive (240 ms)
et la durée maximum d'une pause entre mots.

FIG. 16 - *Automate de détection Bruit/Parole*

Nom : `Segmentation`
Entrées : en, $seuil$

$etat \leftarrow 1$
$ncur \leftarrow 0$
tant que $ncur < lg_{enr}$ faire
  $nsup \leftarrow lg_{enr} - ncur$
  pour $j = 0 \cdots nsup$ faire
    $\mathbf{mu1_{ct}} \leftarrow \mathbf{mu1_{ct}} + (\mathbf{1 - 0.9})(\mathbf{en - mu1_{ct}})$
    $\mathbf{mu3_{ct}} \leftarrow \mathbf{mu3_{ct}} + (\mathbf{1 - 0.9})(\mathbf{en^3 - mu3_{ct}})$
    $\mathbf{sigma_{ct}} \leftarrow \mathbf{sigma_{ct}} + (\mathbf{1 - 0.9})(|\mathbf{en - mu1_{ct}}| - \mathbf{sigma_{ct}})$
    $\mathbf{mom3_{ct}} \leftarrow \mathbf{mu3_{ct}} / \mathbf{sigma_{ct}^3}$
    si $etat = 1$ faire
      $mu1_{lt} \leftarrow mu1_{lt} + (1 - 0.995)(en - mu1_{lt})$
      $\mathbf{mu3_{lt}} \leftarrow \mathbf{mu3_{lt}} + (\mathbf{1 - 0.955})(\mathbf{en^3 - mu3_{lt}})$
      $sigma_{lt} \leftarrow sigma_{lt} + (1 - 0.9)(|en - mu1_{lt}| - sigma_{lt})$
      $\mathbf{mom3_{lt}} \leftarrow \mathbf{mu3_{lt}} / \mathbf{sigma_{lt}^3}$
    fin faire
    $seuil_{detec} \leftarrow mu1_{lt} + sigma_{lt} * seuil$
    $\mathbf{rap} \leftarrow \mathbf{mom3_{ct}} / \mathbf{mom3_{lt}}$
    si $etat = 3$ faire
      $\mathbf{seuil_{rap}} \leftarrow \mathbf{seuil_{rap}} + (\mathbf{1 - 0.99})(\mathbf{coef} * \mathbf{rap} - \mathbf{seuil_{rap}})$
    fin faire

B Description de l'algorithme utilisant les moments d'ordre 3

On présente dans le tableau 8 une partie de l'algorithme de segmentation, utilisant les moments d'ordre 3, non centrés, normalisés, avec un seuil adaptatif. On remarque que le coût de l'algorithme est augmenté par rapport à l'algorithme de base par les estimations à court-terme, les estimations à long-terme des moments d'ordre 3, ainsi que les tests d'hypothèses sur le rapport des moments (apparaissant en gras). Ce coût suplémentaire n'est cependant pas excessivement important.

C Bases de données

Pour la validation des tests il est important d'utiliser plusieurs bases de données. Les tests ont été effectués sur deux bases, l'une enregistrée sur le réseau RTC (cf [Mauuary94]), l'autre sur le réseau GSM (cf [Karray98b]).

début cas :
etat=1
si $en < seuil_{detec}$ faire
$ds \leftarrow ds + 1$
si $ds > sd$ faire
$ds \leftarrow sd$
fin faire
sinon faire
si $\mathbf{rap} > \mathbf{seuil_{rap}}$ faire
$ds \leftarrow ds + 1$
si $ds > sd$ faire
$ds \leftarrow sd$
fin faire
sinon faire
$dp \leftarrow 1$
$etat \leftarrow 2$
fin faire
fin faire
etat=2
si $en < seuil_{detec}$ faire
$ds \leftarrow dp$
$etat \leftarrow 1$
si $ds > sd$ faire
$ds \leftarrow sd$
fin faire
sinon faire
si $\mathbf{rap} > \mathbf{seuil_{rap}}$ faire
$ds \leftarrow dp$
$etat \leftarrow 1$
si $ds > sd$ faire
$ds \leftarrow sd$
fin faire
sinon faire
$dp \leftarrow dp + 1$
$etat \leftarrow 3$
fin faire
fin faire

etat=3
si $en < seuil_{detec}$ faire
$ds \leftarrow 1$
$etat \leftarrow 4$
fin faire
sinon faire
si $\mathbf{rap} > \mathbf{seuil_{rap}}$ faire
$ds \leftarrow 1$
$etat \leftarrow 4$
fin faire
sinon faire
$dp \leftarrow dp + 1$
fin faire
fin faire
fin faire
etat=4
si $en < seuil_{detec}$ faire
$ds \leftarrow ds + 1$
si $ds > sf$ faire
$etat \leftarrow 1$
$ds \leftarrow 0$
$dp \leftarrow 0$
fin faire
sinon faire
si $\mathbf{rap} > \mathbf{seuil_{rap}}$ faire
$ds \leftarrow ds + 1$
$etat \leftarrow 1$
$ds \leftarrow 0$
$dp \leftarrow 0$
fin faire
sinon faire
$dp \leftarrow 1$
$etat \leftarrow 5$
fin faire
fin faire

```
        etat=5
            si en < seuil_detec faire
                ds ← ds + dp
                etat ← 4
                si ds > sf faire
                    etat ← 1
                    ds ← 0
                    dp ← 0
                fin faire
            sinon faire
                si rap > seuil_rap faire
                    ds ← ds + dp
                    etat ← 4
                    si ds > sf faire
                        etat ← 1
                        ds ← 0
                        dp ← 0
                    fin faire
                sinon faire
                    ds ← dp + 1
                    si dp > pm faire
                        etat ← 3
                    fin faire
                fin faire
            fin faire
        fin cas
    fin faire
    ncur ← ncur + nsup
fin faire
```

TAB. 8 - *Algorithme de segmentation.*

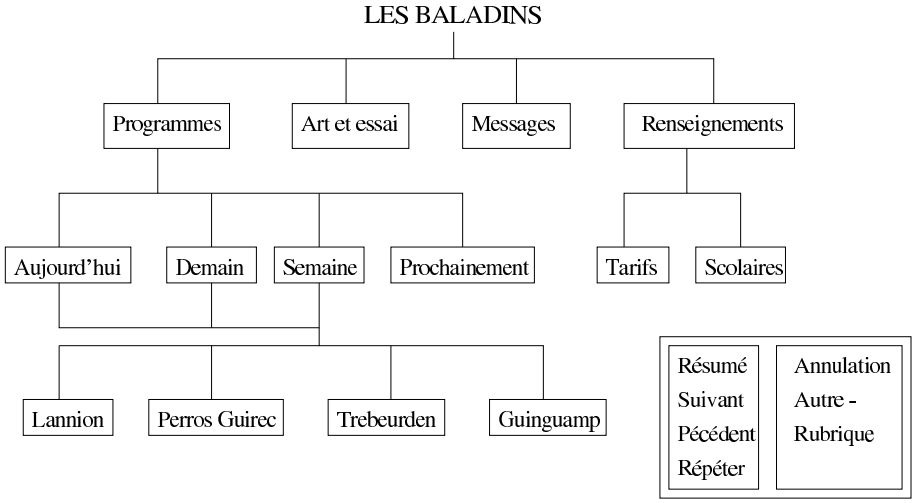


FIG. 17 - *Arborescence du serveur "les Baladins"*

C.1 Les Balladins

La base de données des Balladins est constituée de 1000 appels à un Service Vocal Interactif (SVI) en exploitation qui fournit les programmes de cinéma de la région du Trégor. Les appels enregistrés en continuité sur le réseau RTC ont une durée maximum de 2 min 30 s. Ils contiennent les mots de commande au SVI (soit un vocabulaire de 25 mots), de la parole non destinée au SVI et du bruit. Cette base de données de 1000 appels dure 32 h 25 min. Elle comporte 69.2% de segment de parole, 6.9% de segment d'apparté, le reste étant du bruit, de la parole prononcée par une tierce personne, des raccrochages, et cætera.

C.2 Le corpus GSM

C'est une base de données de laboratoire, où le locuteur (mixte et d'âges variables) répète une liste pré-définie de mots. La base de donnée est constituée de 51 mots : *Autre rubrique, Début, Guide, Pause, Précédent, Reprise, Retour, Sommaire, Stop, Ancien, Annulation, Archivé, Autre choix, Consul-*

tation, Correction, Fin, Information, Message, Messagerie, Nouveau, Répétition, Validation, Annuler, Compléter, Confirmer, Conserver, Consulter, Écouter, Effacer, Enregistrer, Modifier, Quitter, Répéter, Supprimer, Terminer, Valider, Non, Ouais, Oui, Zéro, Un, Deux, Trois, Quatre, Cinq, Six, Sept, Huit, Neuf.

Les appels ont été effectués dans différents types de conditions : à l'intérieur, à l'extérieur, dans un véhicule à l'arrêt ou roulant. Ce corpus contient ainsi différents types de bruit selon l'environnement d'appel. Il est composé de 64.02% de mot du vocabulaire, 6.22% de bruit de fond, 9.28% de bruit GSM (bruits métalliques, trous), 4% d'écho, 2.44% de bruits divers, 6.9% de signal inaudible, et 3.93% de parole hors vocabulaire.

D Système de reconnaissance

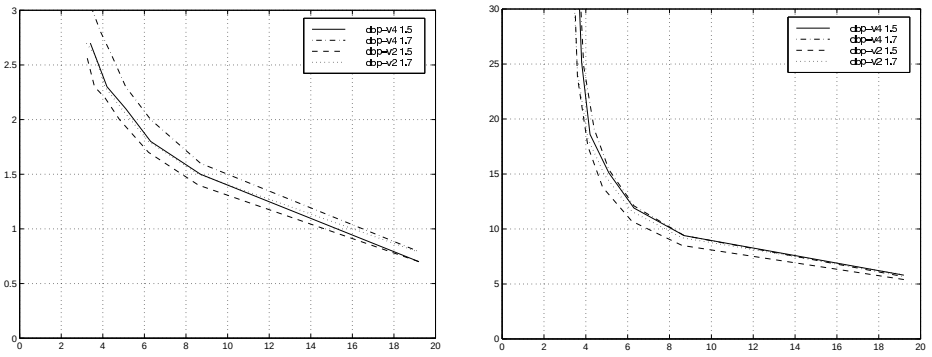
Le système de reconnaissance utilisé pour faire les tests est le logiciel de reconnaissance de la parole PHIL90 développé au CNET. Ce logiciel a été développé pour répondre aux besoins des télécommunications en matière de systèmes de reconnaissance. Sa conception permet de reconnaître une centaine de mots, indépendamment du locuteur. Pour plus de détails on se réfère à [Gagnoulet et al.89]. Les tests ont été effectués selon deux approches.

Tout d'abord un modèle par mots, où l'on considère les mots comme unité de base. Ce modèle s'effectue par un apprentissage, et ne convient donc que pour la description de petits vocabulaires.

Le modèle par allophones (ou flexible) est une approche pour modéliser toutes les réalisations acoustiques possibles d'un phonème. L'unité de base est donc l'allophone. Ce modèle convient mieux que le précédent pour de grands vocabulaires, il est cependant moins performant lorsqu'il y a peu de mots.

E Résultats des tests de reconnaissance par environnement

On présente dans les figures 18, 19, 20, et 21, les résultats de reconnaissance avec un modèle par mots, de l'algorithme initial et celui utilisant les moments d'ordre 3 avec un seuil adaptatif. Le coefficient pour le calcul du seuil adaptatif est fixé à 3. On remarque que les résultats sont sensiblement les mêmes pour les deux algorithmes. L'algorithme utilisant les moments d'ordre 3 donne cependant moins d'erreurs de substitution pour les deux seuils 1.5

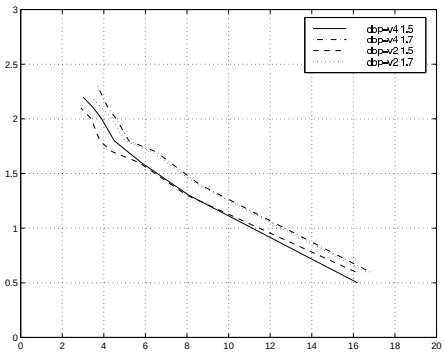


(a) erreurs de substitution (b) fausses alarmes

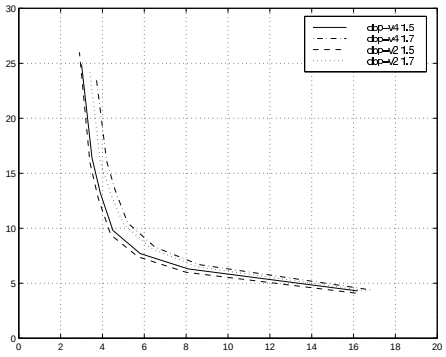
FIG. 18 - *Évaluation par les tests de reconnaissance de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement ; "intérieur"*

et 1.7 étudiés, tandis qu'il donne plus d'erreurs de fausses alarmes. Il n'y a pas de différences flagrantes entre les différents environements.

Les figures 22, 23, 24, et 25, représentent les mêmes résultats des tests de reconnaissance, par environnement, mais avec un modèle flexible. Là encore les courbes en fonction des erreurs de substitution sont proches pour les deux algorithmes, dans tous les environnements. Par contre, on remarque que pour tous les environnements, l'algorithme utilisant les moments d'ordre 3, donne moins d'erreurs de fausses alarmes que l'algorithme initial.

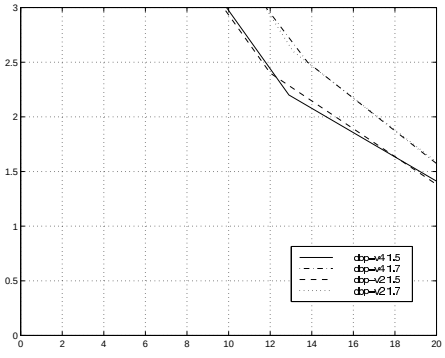


(a) erreurs de substitution

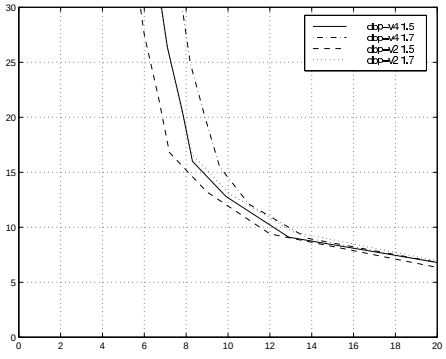


(b) fausses alarmes

FIG. 19 - *Évaluation par les tests de reconnaissance de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement ; "véhicule à l'arrêt"*

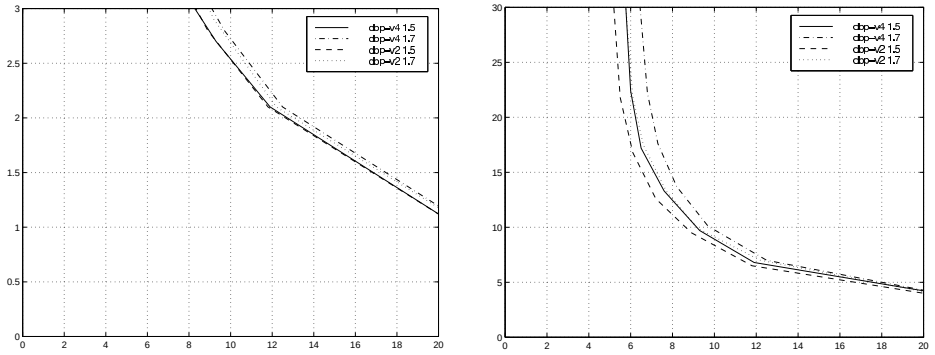


(a) erreurs de substitution



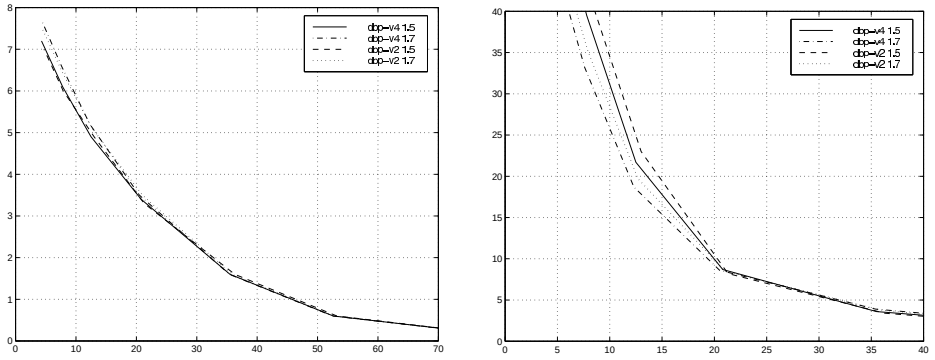
(b) fausses alarmes

FIG. 20 - *Évaluation par les tests de reconnaissance de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement ; "extérieur"*



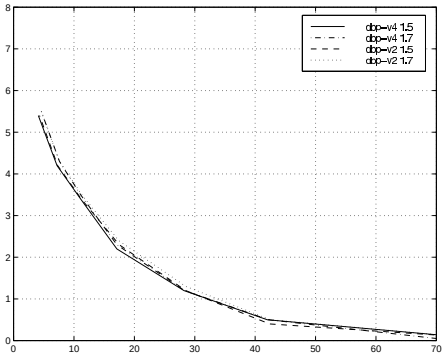
(a) erreurs de substitution (b) fausses alarmes

FIG. 21 - *Évaluation par les tests de reconnaissance de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement ; "véhicule roulant"*

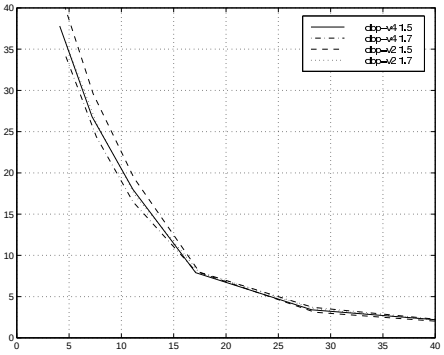


(a) erreurs de substitution (b) fausses alarmes

FIG. 22 - *Évaluation par les tests de reconnaissance, avec modèle flexible, de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement ; "intérieur"*

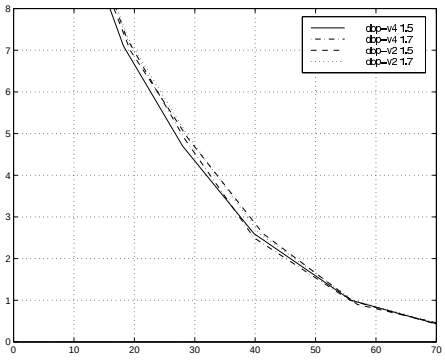


(a) erreurs de substitution

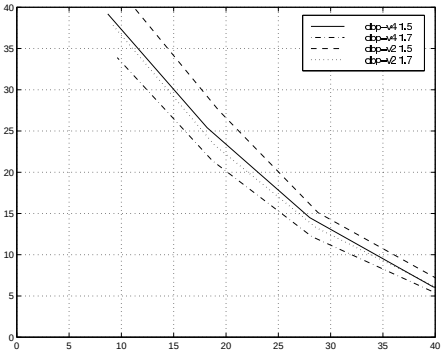


(b) fausses alarmes

FIG. 23 - *Évaluation par les tests de reconnaissance, avec modèle flexible, de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement ; "véhicule à l'arrêt"*

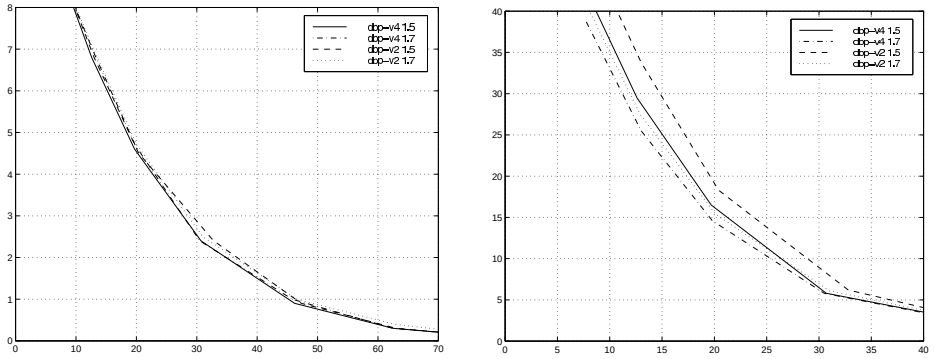


(a) erreurs de substitution



(b) fausses alarmes

FIG. 24 - *Évaluation par les tests de reconnaissance, avec modèle flexible, de l'agorithme adaptatif avec un coefficient fixé à 3, en environnement ; "extérieur"*



(a) erreurs de substitution (b) fausses alarmes

FIG. 25 - *Évaluation par les tests de reconnaissance, avec modèle flexible, de l'agorithme adaptatif avec un coefficient fixé à* 3*, en environnement ; "véhicule roulant"*

Références

[Bouteille99] Bouteille (F.). – *Étude et mise en œuvre de dispositifs de contrôle de l'écho acoustique pour terminaux mains-libres et applications multimédia.* – Diplome de recherche technologique, Université de Rennes 1, 1999.

[Doukas et al.97] Doukas (N.), Naylor (P.) et Stathaki (T). – Voice Activity Detection Using Source Separation Techniques. *European Conference on Speech Communication and Technology*, pp. 1099–1102. – Rhodes, Grèce, septembre 1997.

[Gagnoulet et al.89] Gagnoulet (C.) et Jouvet (D.). – Développement récents en reconnaissance de la parole. *L'Echo des RECHERCHES*, no135, 1989, pp. 27–36.

[Jacovitti et al.91] Jacovitti (G.), Pierucci (P.) et Falaschi (A.). – Speech Segmentation and Classification Using Higer Order Moments. *European Conference on Speech Communication and Technology*, pp. 1371–1374. – Gènes, Italie, septembre 1991.

[Karray98a] Karray (L.). – *Estimation des Statistiques du Bruit et de la Parole pour une Détection Bruit/Parole plus Robuste.* – Rapport technique n° 8, DT/DIH/DIPS/285, avril 1998.

[Karray98b] Karray (L.). – *Nouveau Critère pour l'Automate de Détection Bruit/Parole.* – Rapport technique n° 3, DT/DIH/DIPS/48, janvier 1998.

[Lacoume et al.97] Lacoume (J.-L.), Amblard (P.-O.) et Comon (P.). – *Statistiques d'ordre supérieur pour le traitement du signal.* – Masson, 1997.

[Mauuary94] Mauuary (L.). – *Amélioration des performances des serveurs vocaux interactifs.* – Thèse de Doctorat, Université de Rennes 1, 1994.

[Mccullagh87] McCullagh (P.). – *Tensor Methods in Statistics.* – Charpman and Hall, 1987.

[Pincibono93] Pincibono (B.). – *Signaux aléatoires*. – Dunod Université, 1993.