Utilisation des moments d'ordre 3 pour une détection Parole/non-Parole robuste

Arnaud Martin

France Télécom R&D/DIH/DIPS
2, avenue Pierre Marzin – 22307 Lannion, France
Tél.: ++33 (0)296 05 23 10 - Fax: ++33 (0)296 05 35 30
Mél: arnaud.martin@rd.francetelecom.fr

ABSTRACT

In noisy environment, robustness to noise of speech/non-speech detection is necessary for speech recognition. This paper presents a new method for speech/non-speech detection using third-order statistics. The analysis of the energy third-order statistic behavior gives useful information on the energy distribution. The new algorithm is evaluated in terms of segmentation and recognition performance. Different telephone call environments are considered for the evaluation. This algorithm is compared to the one based on noise and speech statistics presented in [Kar98a]. The results show that the new algorithm outperforms the one based on second order noise and speech statistics only, especially in the case of noisy environment.

1. INTRODUCTION

Les performances de la reconnaissance vocale décroissent en utilisation dans des environnements très bruités. Une détection efficace des périodes de parole et de non-parole est cruciale pour la reconnaissance.

C'est dans ce sens que de nombreuses études ont été menées. La caractéristique principale du signal utilisée est l'énergie, mais on peut lui associer par exemple la fréquence fondamentale [Iwa99] pour détection plus fine. Un algorithme de détection de parole/non-parole fondé sur l'utilisation de l'estimation des statistiques de l'énergie des périodes parole et des périodes de non-parole, a été présenté dans [Kar98]. Cet algorithme nous servira d'algorithme initial. Des méthodes comme la logique floue peuvent également être utilisées [Ber99].

L'observation que la distribution du signal de parole est non gaussienne a conduit différentes études à considérer les statistiques d'ordre supérieur dans des systèmes de détection de la parole. Dans [Jac91], il est proposé l'utilisation du skewness et du kurtosis (qui sont les cumulants d'ordre 3 et 4 normalisés). [Dou97] utilise le fait que le cumulant croisé de deux variables aléatoires indépendantes soit nul, pour discriminer le signal parole et celui du bruit à la source. Dans [Nem99], les auteurs intègrent le cumulant d'ordre 4 du résidu des coefficients de prédiction LPC dans un système de détection d'activité vocale. On se propose ici, d'intégrer le moment d'ordre 3, normalisé, conditionnellement à l'algorithme initial utilisant les statistiques de l'énergie dans des périodes de parole et de non-parole. Cette statistique permet une description plus fine de la distribution de l'énergie qui fournit la décision dans le système de détection Parole/non-Parole.

Ce papier est organisé comme suit. Dans la section 2, on rappellera l'algorithme initial de détection parole/non-parole, utilisant les statistiques de l'énergie des périodes de parole et de non-parole. Dans la section 3, est présentée l'estimation des moments d'ordre supérieur, et en particulier du moment d'ordre 3. On introduit ensuite un nouveau critère de détection de parole et de non-parole fondé sur le moment d'ordre 3, que l'on intègrera dans l'algorithme initial. Dans la section 4, on présente les performances de cet algorithme de détection en comparaison avec l'algorithme initial. Les évaluations sont faites sur deux bases de données téléphoniques, enregistrées l'une à travers le réseau RTC, l'autre à travers le réseau GSM, dans différents environnements d'appel.

2. ALGORITHME INITIAL

On rappelle dans cette section l'algorithme de détection de parole/non-parole, fondé sur un automate à cinq états [Mau94]. Les cinq états de l'automate sont : *silence, présomption de parole, parole, plosive ou silence et reprise possible de la parole.* Les transitions d'un état à l'autre se font à l'aide de contraintes de durée, et par des tests sur les caractéristiques du signal. Ces différents tests correspondent à différents critères.

Les transitions d'un état à l'autre de l'automate se font par un test d'hypothèse sur chaque trame du signal observé. On considère la moyenne et la variance de l'énergie dans les périodes de parole et dans celles de non-parole [Kar98]. L'estimation des statistiques de l'énergie des périodes de parole se fait dans l'état *parole*, celles des périodes de non-parole dans l'état s*ilence*. Cette approche découle d'une approche Bayésienne. On teste deux hypothèses :

H_0 : on est dans un état de non-parole,

H_1 : on est dans un état de parole.

On considère dans un premier temps que les distributions de l'énergie dans les périodes de parole et

dans celles de non-parole suivent deux lois normales. La décision du passage d'un état à l'autre de l'automate se fera, pour chaque trame observée x par comparaison du maximum de vraisemblance $P(H_i / x)$ de chaque hypothèse, pour i=0 ou 1. En supposant les deux hypothèses équiprobables, et utilisant la formule de Bayes, on se ramène à la comparaison à un seuil du rapport de vraisemblance :

$$R(x) = \frac{P(x / H_0)}{P(x / H_1)} .$$

3. CRITERE DU MOMENT D'ORDRE 3

L'étude expérimentale du rapport du moment d'ordre 3 dans les périodes de parole et de bruit montre que c'est un paramètre pertinent pour affiner la description de la distribution de l'énergie du bruit et de la parole. La décision faite uniquement sur la moyenne et l'écart type de l'énergie va donc pouvoir être précisée.

3.1 Estimation des statistiques d'ordre supérieur

L'estimation ''classique'' des moments d'ordre *r*, $\hat{\mu}_r$ de l'énergie est l'estimation arithmétique :

$$\hat{\mu}_r(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^r$$

où x_i est l'énergie du signal pour la $i^{ème}$ trame, et N est le nombre de trames. Cette estimation a l'inconvénient de ne pas tenir compte de la non-stationnarité du signal de parole. On utilise donc ici une estimation sur des fenêtres exponentielles, qui revient à pondérer les trames avec des poids décroissant avec le temps. Ainsi pour une trame donnée n, le moment d'ordre r $\hat{\mu}_r$ est définit par :

$$\hat{\mu}_r(n) = \hat{\mu}_r(n-1) + (1-\lambda)(x_n^r - \hat{\mu}_r(n-1))$$

où λ est le facteur d'oubli. Le degré supposé de stationnarité du signal détermine le facteur λ, et ainsi le nombre de trames qui seront considérées pour le calcul de la statistique. Cet estimateur contrairement à l'estimateur arithmétique n'est pas sans biais, en effet on a :

$$E[\hat{\mu}_r(n)] = (1-\lambda^{n+1})\mu_r$$

où μ_r est le moment d'ordre r théorique. Cet estimateur est asymptotiquement sans biais. L'étude de la variance de ces estimateurs reste un problème délicat quant à l'établissement des formules générales. Une étude de ces estimateurs pour les grandes valeurs de n a été réalisée dans [McC87]. Il est montré que ces estimateurs sont consistants, avec une vitesse de convergence diminuant avec la croissance de l'ordre du moment.

3.2 Le moment d'ordre 3

On considère ici le moment d'ordre 3, normalisé, définit par :

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

où $\hat{\mu}_3$ et $\hat{\sigma}$ sont respectivement l'estimateur du moment d'ordre 3 et de l'écart type de l'énergie. Ces estimateurs sont calculés de la façon décrite dans la section 3.1, sur des fenêtres exponentielles. Le moment d'ordre 3, normalisé, d'une quantité de moyenne nulle est exactement le skewness. De plus, la variance de cet estimateur reste suffisamment faible à l'échelle de l'hypothèse de stationnarité du signal, elle est notamment plus faible que celle du skewness et celle du moment d'ordre 3 centré.

3.3 Intégration du moment d'ordre 3

Nous allons voir comment intégrer ce paramètre dans l'algorithme initial. $\hat{m}_3$ est calculé avec deux facteurs d'oubli différents : l'un ($\lambda_{ct} = 0.9$) qui donnera une estimation à court terme du moment d'ordre 3, l'autre ($\lambda_{lt} = 0.99$) qui donnera une estimation à long terme. L'estimation à long terme est calculée récursivement uniquement dans les périodes de non-parole, c'est-à-dire dans l'état *silence* de l'automate. La décision sera prise en comparant le rapport $rap(n) = \frac{\hat{m}_{3ct}(n)}{\hat{m}_{3lt}(n)}$, des estimations à court terme et à long terme du moment d'ordre 3, à un seuil adaptatif. Ce rapport est plus faible dans les périodes de parole. Cette décision est prise conditionnellement au test sur l'énergie de l'algorithme initial, et aux contraintes de temps. C'est-à-dire que pour le passage d'un état à l'autre, on comparera d'abord le rapport $R(x)$ à un seuil, la décision sera confirmée par le test du moment d'ordre 3. Le seuil adaptatif est calculé récursivement dans les périodes détectées comme de la parole par l'algorithme initial, de la façon suivante :

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

où λ_T est un facteur d'oubli, et $c > 1$ est un coefficient permettant d'obtenir une borne supérieure du rapport des moments dans les périodes de parole. Il a été optimisé à une valeur proche de 3 sur nos bases de données

Cette méthode suppose deux hypothèses, les distributions énergétiques du bruit et de la parole ont des moments d'ordre 3 différents, et le bruit est plus stationnaire que la parole.

4. EXPERIMENTATIONS

Les tests ont été effectués sur deux bases de données. Pour évaluer le nouvel algorithme on a procédé à une évaluation de la segmentation et à une évaluation de la

reconnaissance à partir d'un système de reconnaissance élaboré au CNET [Mok97].

4.1 Bases de données

Une première base de données est constituée de 1000 appels téléphoniques à un serveur vocal interactif en exploitation, donnant les programmes de cinéma. Les appels enregistrés en continuité à travers le réseau RTC contiennent les mots de commande au serveur (soit un vocabulaire de 25 mots). Le corpus obtenu par la segmentation manuelle contient 67% de mots du vocabulaire, 11% de parole hors vocabulaire et 22% de bruit.

La deuxième base de données est une base enregistrée par téléphone (hors contexte applicatif), constituée de 51 mots de vocabulaire que chaque locuteur répète. Les 395 appels ont été effectués à travers le réseau GSM, à partir de différents environnements (*intérieur, extérieur, voiture à l'arrêt, voiture roulant*). Le corpus a été segmenté manuellement, 68% des segments sont des mots du vocabulaire, 4% des mots hors vocabulaire et 28% des bruits.

4.2 Test de segmentation

Les tests de segmentation sont effectués par comparaison à la base segmentée manuellement. Les segments de parole du vocabulaire, hors vocabulaire et différents types de bruits ont été annotés. Ainsi différentes erreurs apparaissent, les omissions les insertions, les regroupements et les fragmentations [Mau94]. En vue de la reconnaissance ces erreurs sont classées en erreurs rejetables (comportant les insertions et les détections de parole hors vocabulaire) et en erreurs définitives (comportant les omissions, les fragmentations et les regroupements). Les courbes sont obtenues en faisant varier le seuil de détection de l'algorithme initial (seuils indiqués sur les courbes).

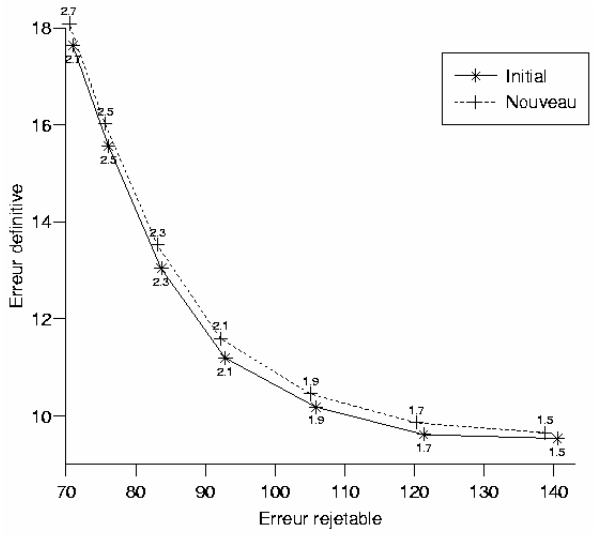


Figure1: Test de segmentation – données RTC

La figure 1 présente les résultats des tests sur la base de données enregistrée à travers le réseau RTC. Dans cet environnement non bruité, l'algorithme initial présente des points de fonctionnement conduisant à moins d'erreurs pour la reconnaissance. On remarque que pour un même seuil le nouvel algorithme donne davantage d'erreurs définitives (dû aux fragmentations), mais un peu moins d'erreurs rejetables (dû aux insertions).

La figure 2 donne les résultats des tests sur la base de données enregistrées à travers le réseau GSM. On a regroupé ici les quatre environnements. Le nouvel algorithme présente un peu moins d'erreurs en vue de la reconnaissance. Lorsque les tests sont effectués séparément sur chaque environnement l'écart est plus prononcé pour les environnements les plus bruités (*extérieur, véhicule roulant*).

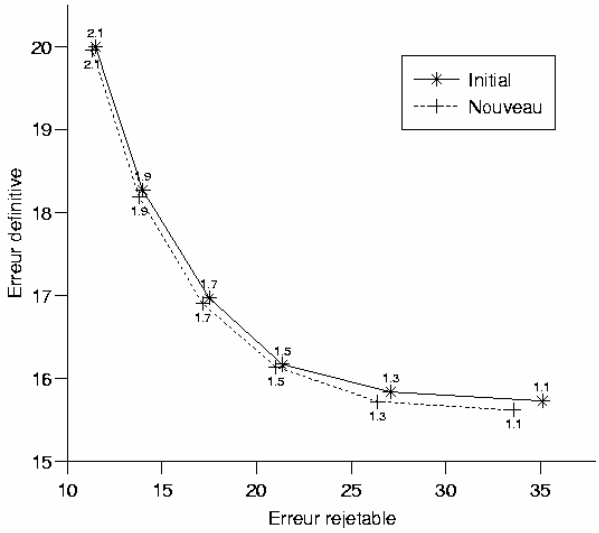


Figure2: Test de segmentation – données GSM

4.3 Test de reconnaissance

Le système de reconnaissance utilisé au CNET [Mok97] est fondé sur la modélisation des mots du vocabulaire à partir des chaînes de Markov. Pour modéliser toutes les réalisations acoustiques possibles, on utilise un modèle par allophones (modélisation contextuelle des phonèmes). Cette modélisation est faite à partir d'une base différente de celles utilisées pour les tests. Les courbes des tests de reconnaissance sont obtenues en faisant varier l'importance du rejet. On représente ici les erreurs de substitution associées aux fausses acceptations en fonction des faux rejets. Pour chaque algorithme deux seuils ont été choisis pour la détection.

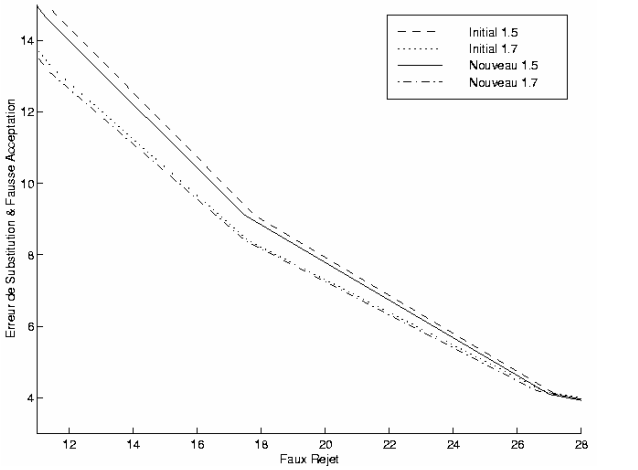
Les résultats des tests sur la base enregistrée à travers le réseau RTC sont équivalents pour les deux algorithmes. Le peu d'erreurs définitives en plus est compensé par moins d'erreurs rejetables.

La figure 3 donne les résultats des tests sur la base de données enregistrée à travers le réseau GSM. Le

nouvel algorithme présente des taux d'erreur légèrement plus faibles que l'algorithme initial. La différence se situe surtout au niveau des fausses acceptations. Ceci vient du fait qu'il y a moins d'erreurs rejetables. La figure 4 montre que ces résultats sont encore plus accentués dans le cas d'un environnement bruité où les enregistrements ont été effectués lors dans un véhicule roulant.

4.4 Discussion

Les résultats obtenus montrent une légère amélioration du nouvel algorithme dans des conditions bruitées. Ceci vient du fait que le nouveau critère a été utilisé conditionnellement au critère de l'algorithme initial. Cependant cette amélioration reste peu significative. En effet l'écart des taux d'erreur reste très faible. La diminution des erreurs se retrouve surtout au niveau des erreurs rejetables. Les omissions et les fragmentations ne sont pas diminuées (les erreurs définitives). Cette approche diminue cependant les fragmentations dans des environnements très bruités.



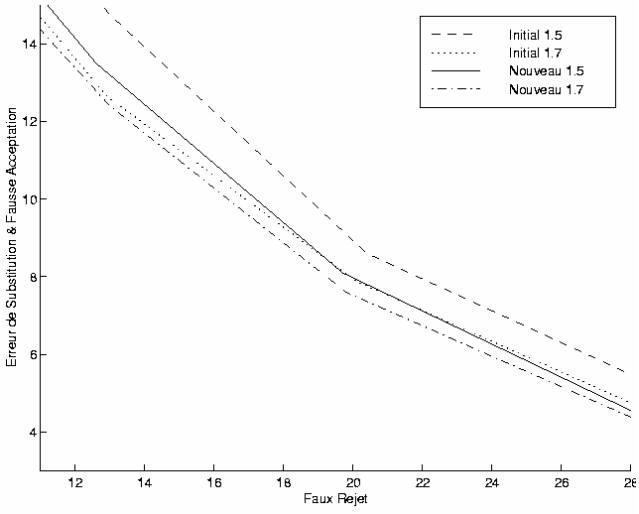


Figure3: Test de reconnaissance – données GSM

Figure4: Test de reconnaissance – données GSM véhicule roulant

5. CONCLUSION ET PERSPECTIVES

Nous avons intégré un critère sur les moments d'ordre 3 de l'énergie du signal dans l'algorithme de détection de parole/non-parole utilisant les statistiques des périodes de parole et de non-parole. Le rapport des moments d'ordre 3 à court terme et à long terme nous a permis de comparer les distributions de l'énergie des périodes de parole et de non-parole. Les tests de segmentation et de reconnaissance ont montré une légère amélioration de l'algorithme utilisant ce nouveau critère. Cette amélioration reste cependant peu significative.

Nous avons ici introduit les moments d'ordre 3 conditionnellement à la décision prise par l'algorithme initial, dans l'état de *parole* de l'automate. Ce nouveau critère pourrait conduire à une autre approche. Nous nous proposons dans une prochaine étude de combiner une décision fournit par ce nouveau critère avec la décision prise par l'algorithme initial, par le biais des méthodes de fusion de décision.

BIBLIOGRAPHIE

[Ber99] Beritelli F., Casale S. and Cavallaro K. (1999), "A Multi-Channel Speech/Silence Detector based on Time Delay Estimation and Fuzzy Classification", ICASSP, Vol. 1, pp. 93- 96.

[Iwa99] Iwano K. and Hirose K. (1999), "Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition", ICASSP, Vol. 1, pp. 133- 136.

[Jac91] Jacovitti G., Pierucci P. and Falashi A.(1991), "Speech Segmentation and Classification Using Higher Order Moments", Eurospeech, pp. 1371-1374.

[Dou97] Doukas N., Naylor P. and Stathaki T. (1997), "Voice Activity Dection Using Source Separation Techniques", Eurospeech, pp. 1099-1102.

[Kar98] Karray L. and Monné J. (1998) "Robust speech/non-speech detection in adverse conditions based on noise and speech statistics", ICSLP, Vol.. 4, pp. 1471-1474.

[Mau94] Mauuary L. (1994) "Amélioration des serveurs vocaux interactifs", Thèse de Doctorat, université de Rennes 1.

[McC87] McCullagh P. (1987) "Tensor Methods in Statistics", Charpman and Hall.

[Mok97] Mokbel C. *et al*. (1997) "Towards improving ASR robustness for PSN and GSM telephone applications", Speech communication, Vol. 23, pp 141-159.

[Nem99] Nemer E., Gourbran R. and Mahmoud S.

(1999) “The fourth-order cumulant of speech signals with application to voice activity detection”, Eurospeech, Vol. 5, pp. 2391-2394.