Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Comparaison de critères de segmentation par détection de ruptures sur un signal sonore

Mouhamadou SECK, Frédéric BIMBOT, Bernard DELYON

IRISA - SIGMA2 / C.N.R.S & INRIA
Campus Universitaire de Beaulieu, 35042 RENNES Cedex, FRANCE
e-mail : mseck@irisa.fr, bimbot@irisa.fr, delyon@irisa.fr

Résumé – Cet article présente une méthode de segmentation de signaux, par détection de ruptures. Cette méthode consiste à calculer à chaque instant un critère de présence de rupture, basé sur une statistique de test d'existence d'une rupture. La décision de présence ou non d'une rupture est prise elle aussi à chaque instant, en comparant la valeur du critère à un seuil. Deux statistiques de détection de rupture, le rapport de vraisemblance généralisé et la statistique de la divergence, sont évaluées sur un signal constitué de réels segments de parole et de musique en alternance. Les performances obtenues avec ces deux statistiques ne sont pas significativement différentes et tournent autour de 5% de taux d'erreurs, quand le taux de non-détections est égal au taux de fausses alarmes. Toutefois, le rapport de vraisemblance généralisé a une plus grande tendance à la non-détection, tandis que la statistique de la divergence présente une plus grande tendance aux fausses alarmes.

Abstract – This article presents a technique for discrete time signal segmentation, using a change point detection approach. This technique consist of computing at each time, a change point criterion based on a change point detection statistic. Change point decision is taken a each time, by comparing the criterion value with a threshlod. Two change point statistics, the generalized likelihood ratio and the divergence statistic, are evaluated on a signal with real speech and music segments. Performances obtained with these two statistics, are not significantly different and equal error rates are around 5%. However, the generalized likelihood ratio has a tendency to more non-detections, while the divergence statistic is more inclined to false alarms.

Introduction

La segmentation est un prétraitement courant pour de nombreuses applications en Traitement du Signal (classification, indexation, surveillance...). Cette opération consiste à découper un signal en intervalles de temps (segments) pendant lesquels, certaines de ses caractéristiques restent stationnaires. Par exemple, dans le cas d'un signal audio, les segments peuvent correspondre à des classes de sons comme la parole, la musique, la parole d'un locuteur, etc... Les distributions des caractéristiques du signal au niveau des segments peuvent être connues ou inconnues. Dans cet article, nous nous plaçons dans le cas où les distributions au niveau des segments sont inconnues.
La section 1 présente le principe de la méthode de segmentation proposée. La section 2 donne une formulation du problème et la section 3 décrit de manière détaillée la méthode de segmentation proposée. Dans la section 4, nous présentons deux statistiques de test d'existence d'une rupture, qui sont à la base des indices de rupture que nous utilisons. Des expériences en segmentation parole/musique, où sont utilisés les deux indices de rupture, sont présentées dans la section 5 et les résultats sont discutés dans la section 6. La dernière section est consacrée à la conclusion et aux perspectives.

1 Méthode proposée

Le signal est d'abord découpé en trames régulières. Chacune des trames est représentée par un vecteur de paramètres acoustiques. Pour chaque trame (instant), nous calculons un indice de présence de rupture (positif). Les indices que nous comparons ici, sont basés sur une dissimilarité (au sens large du terme) au voisinage de la trame, entre le modèle ajusté au passé de la trame et le modèle ajusté à son futur. De cet indice, est extrait un critère de présence de rupture (qui est nul aux instants qui ne sont pas maximum local de l'indice). La décision est prise trame à trame, par comparaison du critère à un seuil.
Dans nos expériences, nous utilisons comme paramètres acoustiques, les Composantes Principales du logarithme du module du Spectre (CPS)[1]. Les modèles ajustés au passé et au futur de chaque instant sont des modèles Gaussiens.

2 Formulation du problème

Soit $\mathcal{Y}_1^n = \{Y_1, \cdots, Y_n\}$ une séquence d'observations d'un signal à temps discret, à valeurs dans $\mathbb{R}^d$. Les observations correspondent à une séquence de vecteurs de paramètres acoustiques, extraits d'un signal sonore. On utilisera la notation $\mathcal{Y}_i^j$, pour désigner la séquence d'observations $\{Y_i, \cdots, Y_j\}$.
On suppose qu'il existe un entier k et des instants non observés $r_1^\star < r_2^\star < \cdots < r_k^\star$, tels que le signal est stationnaire sur chaque intervalle de temps $[r_l^\star + 1, r_{l+1}^\star]$ (avec la

1. Les facteurs principaux sont estimées sur une base d'apprentissage. Les CPS sont utilisées pour la classification parole/musique de segments sonores dans [SBD98]

convention $r_0^\star = 0$ et $r_{k+1}^\star = n$). Les instants $r_l^\star$ sont des instants de changement de distribution (stationnaire) du signal et sont appelés instants de rupture. On souhaite localiser ces instants dans la séquence d'observations $\mathcal{Y}_1^n$. Ce problème est connu sous le nom de détection-estimation de ruptures. Dans cette étude, nous nous plaçons dans le cas où :

- les observations sont supposées statistiquement indépendantes;
- les distributions du signal à l'intérieur des segments sont décrites par une famille paramétrée de densités $\{p(.|\theta); \theta \in \Theta \subset \mathbb{R}^m\}$;
- les instants de ruptures sont supposés déterministes.

3 Description de la méthode

La méthode de segmentation que nous proposons, consiste d'abord à calculer un indice de rupture, à chaque instant t. On note par $S(t)$ la valeur de cet indice à l'instant t, pour $t = 1, \cdots, n$. À partir de l'indice S, un critère de présence de rupture, également défini pour chaque instant, est calculé. $C(t)$ désigne la valeur du critère à l'instant t. La décision de présence d'une rupture est prise à chaque instant, par comparaison de la valeur critère $C(t)$ à un seuil. La section 3.1 présente les deux indices que nous comparons et la section 3.2 décrit la manière dont nous obtenons le critère de décision.

3.1 Indices de rupture

Les deux indices de rupture que nous étudions correspondent au calcul, en un instant donné t, d'une dissimilarité (au sens large du terme) entre le passé $\mathcal{Y}_{t-L+1}^t$ et le futur $\mathcal{Y}_{t+1}^{t+L}$, où L représente la taille commune du passé et du futur. Ces deux indices sont :

$$S_1(t) = \frac{1}{L} \log \frac{p(\mathcal{Y}_{t-L+1}^t|\hat{\theta}_1)\, p(\mathcal{Y}_{t+1}^{t+L}|\hat{\theta}_2)}{p(\mathcal{Y}_{t-L+1}^{t+L}|\hat{\theta}_0)}$$

$$S_2(t) = \frac{1}{L}\left[\log \frac{p(\mathcal{Y}_{t-L+1}^t|\hat{\theta}_1)}{p(\mathcal{Y}_{t-L+1}^t|\hat{\theta}_2)} + \log \frac{p(\mathcal{Y}_{t+1}^{t+L}|\hat{\theta}_2)}{p(\mathcal{Y}_{t+1}^{t+L}|\hat{\theta}_1)}\right]$$

où :

$$\begin{cases} \hat{\theta}_1 = \arg\max_{\theta\in\Theta} p(\mathcal{Y}_{t-L+1}^t|\theta) \quad \hat{\theta}_2 = \arg\max_{\theta\in\Theta} p(\mathcal{Y}_{t+1}^{t+L}|\theta) \\ \hat{\theta}_0 = \arg\max_{\theta\in\Theta} p(\mathcal{Y}_{t-L+1}^{t+L}|\theta) \end{cases}$$

Cette approche par comparaison entre le passé et le futur proches de l'instant courant t, est très utilisée en segmentation de signaux [Bra83, DAOS88]. L'indice $S_1(t)$ correspond, au facteur multiplicatif $\frac{1}{L}$ près, au logarithme du rapport de vraisemblance généralisé (RVG). Quant à $S_2(t)$, c'est une approximation de la divergence de Kullback [2] entre le modèle ajusté au passé $\mathcal{Y}_{t-L+1}^t$ et celui ajusté au futur $\mathcal{Y}_{t+1}^{t+L}$. L'indice S_2 présente l'avantage de n'utiliser (estimer) que deux modèles ($\hat{\theta}_1$ et $\hat{\theta}_2$), au lieu des trois modèles ($\hat{\theta}_0$, $\hat{\theta}_1$ et $\hat{\theta}_2$) utilisés par S_1.

2. La divergence de Kullback est une quantité positive qui mesure la dissemblance entre deux distributions.

3.2 Critère de décision

La décision de présence ou non d'une rupture, ne peut pas être prise en comparant directement l'indice $S(t)$ à un seuil, à cause du risque de fausses alarmes, notamment au voisinage des maxima de l'indice. Nous utilisons l'approche suivante : pour un instant t donné, nous recherchons les premiers instants $\tau_1(t) < t$ et $\tau_2(t) > t$, tels que :

$$S(\tau_1(t)) > S(t) \;\; et \;\; S(\tau_2(t)) > S(t)$$

Puis nous calculons les quantités $v_1(t)$ et $v_2(t)$:

$$v_1(t) = \min_{\tau_1(t)<i<t} S(i) \;\; et \;\; v_2(t) = \min_{t<i<\tau_2(t)} S(i)$$

Et enfin : $u(t) = \max\{v_1(t), v_2(t)\}$

Le critère de présence de rupture à l'instant t est alors défini par : $C(t) = S(t) - u(t)$

Le critère C ainsi défini, est nul aux instants qui ne sont pas maximum local de l'indice S. La décision de présence où non d'une rupture à l'instant t, est prise en comparant la valeur du critère à un seuil :

$$C(t) \underset{NONRUPT}{\overset{RUPT}{\gtrless}} \lambda$$

4 Test d'existence d'une rupture : statistiques de base des indices

Les deux indices de rupture S_1 et S_2, définis à la section 3.1, ont été obtenus à partir de statistiques du test d'existence d'une rupture sur une séquence d'observations $\mathcal{X}_1^n = \{X_1, \cdots, X_n\}$ dans $\mathbb{R}^d$, statistiquement indépendantes. La famille de distributions reste celle définie à la section 2. On considère alors le test basé sur n observations, de l'hypothèse d'absence de rupture $H_0(n)$ contre l'hypothèse de présence de rupture $H_1(n)$.

- $H_0(n)$: *$\{X_1, \cdots, X_n\}$ suivent la même distribution.*
- $H_1(n)$: *Il existe un instant $r^\star(n) \geq 1$, tel que $\{X_1, \cdots, X_{r^\star(n)}\}$ suivent la distribution de paramètre $\theta_1^\star$ et $\{X_{r^\star(n)+1}, \cdots, X_n\}$ suivent la distribution de paramètre $\theta_2^\star \neq \theta_1^\star$.*

Les sections 4.1 et 4.2 présentent respectivement les statistiques de test d'existence d'une rupture qui sont à la base des indices S_1 et S_2, et quelques unes de leurs propriétés. On introduit les notations suivantes qui seront utilisées dans toute la suite de cette section 4. On note par $\hat{\theta}_1$, $\hat{\theta}_2$ et $\hat{\theta}_0$, les paramètres de distribution définis par :

$$\begin{cases} \hat{\theta}_1 = \arg\max_{\theta\in\Theta} p(\mathcal{X}_1^r|\theta) \quad \hat{\theta}_2 = \arg\max_{\theta\in\Theta} p(\mathcal{X}_{r+1}^n|\theta) \\ \hat{\theta}_0 = \arg\max_{\theta\in\Theta} p(\mathcal{X}_1^n|\theta) \end{cases}$$

Les paramètres $\hat{\theta}_1$ et $\hat{\theta}_2$ dépendent de l'instant r.

4.1 Rapport de vraisemblance généralisé

Pour le test d'existence d'une rupture, une des statistiques de test les plus utilisées est le Rapport de Vraisemblance Généralisé (RVG). La statistique du RVG pour le test d'existence d'une rupture, notée $R(n)$, est définie par :

$$R(n) = \sup_{r=m,\cdots,n-m} \frac{p(\mathcal{X}_1^r|\hat{\theta}_1)\, p(\mathcal{X}_{r+1}^n|\hat{\theta}_2)}{p(\mathcal{X}_1^n|\hat{\theta}_0)}$$

Une étude du comportement asymptotique de la statistique $R(n)$, sous la suite l'hypothèses d'absence de rupture $H_0(n)$, est présentée dans [DP86]. Sous $H_0(n)$ et certaines hypothèses de régularité sur la famille de distributions, Deshayes et Picard montrent que la statistique

$$\Lambda_1(n) = \sup_{r=m,\cdots,n-m} \psi_1(\frac{r}{n}) \log \left[\frac{p(\mathcal{X}_1^r|\hat{\theta}_1)\, p(\mathcal{X}_{r+1}^n|\hat{\theta}_2)}{p(\mathcal{X}_1^n|\hat{\theta_0})} \right]$$

converge en loi vers $\Lambda_1 = \sup_{\eta\in[0,1]} \frac{\psi_1(\eta)}{2\eta(1-\eta)} B_m^T(\eta) B_m(\eta)$.
Dans cette relation, ψ_1 est une fonction définie sur $[0,1]$ telle que $\int_0^1 \left[\frac{\psi_1(\eta)}{\eta(1-\eta)}\right]^2 d\eta < +\infty$, et B_m est un pont Brownien de dimension m, dont les composantes sont indépendantes. Ce résultat met en évidence la nécessité de pondérer la statistique du RVG avant le supremum sur tous les instants r.

4.2 Statistique de la divergence

En présence de rupture, le fait que le modèle $\hat{\theta}_0$ soit estimé sur des observations issues d'un mélange de deux distributions, ne facilite pas l'étude du comportement des statistiques $R(n)$ et $\Lambda_1(n)$. Nous proposons une autre statistique de test d'existence d'une rupture, notée $\Lambda_2(n)$.

$$\Lambda_2(n) = \sup_{r=m,\cdots,n-m} \psi_2(\frac{r}{n}) \log \left[\frac{p(\mathcal{X}_1^r|\hat{\theta}_1)}{p(\mathcal{X}_1^r|\hat{\theta}_2)} \frac{p(\mathcal{X}_{r+1}^n|\hat{\theta}_2)}{p(\mathcal{X}_{r+1}^n|\hat{\theta}_1)} \right]$$

où: ψ_2 est une fonction définie sur $[0,1]$ telle que $\int_0^1 \left[\frac{\psi_2(\eta)}{\eta^2(1-\eta)^2}\right]^2 d\eta < +\infty$.
Soit $\theta_0^\star$ le vrai paramètre de distribution des observations $\mathcal{X}_1^n$ en absence de rupture, $\eta \in]0,1[$, $\hat{\delta}_1$, $\hat{\delta}_2$ et r tels que:

$$\begin{cases} r = [\eta n] \\ \hat{\delta}_1 = \arg\max_{\delta_1} \lim_{n\to+\infty} p(\mathcal{X}_1^r|\theta_0^\star + \frac{\delta_1}{\sqrt{n}}) \\ \hat{\delta}_2 = \arg\max_{\delta_2} \lim_{n\to+\infty} p(\mathcal{X}_{r+1}^n|\theta_0^\star + \frac{\delta_2}{\sqrt{n}}) \end{cases}$$

Sous $H_0(n)$ et les mêmes hypothèses de régularité sur la famille de distributions que pour Λ_1, on vérifie que:

$$\lim_{n\to+\infty} \log \left[\frac{p(\mathcal{X}_1^r|\theta_0^\star + \frac{\hat{\delta}_1}{\sqrt{n}})}{p(\mathcal{X}_1^r|\theta_0^\star + \frac{\hat{\delta}_2}{\sqrt{n}})} \frac{p(\mathcal{X}_{r+1}^n|\theta_0^\star + \frac{\hat{\delta}_2}{\sqrt{n}})}{p(\mathcal{X}_{r+1}^n|\theta_0^\star + \frac{\hat{\delta}_1}{\sqrt{n}})} \right] = \frac{B_m^T(\eta) B_m(\eta)}{2\eta^2(1-\eta)^2}$$

Ce résulat ne signifie pas en soi, une convergence en loi de $\Lambda_2(n)$ vers $\Lambda_2 = \sup_{\eta\in[0,1]} \frac{\psi_2(\eta)}{2\eta^2(1-\eta)^2} B_m^T(\eta) B_m(\eta)$, mais il conforte l'hypothèse d'un résultat de convergence sous $H_0(n)$ du même type que celui énoncé pour $\Lambda_1(n)$.
Nous allons énoncer une propriété illustrative de l'indice de la divergence S_2 dans le cas de modèles Gaussiens.

Propriété 1 *Soit $\tilde{S}_2$ la statistique définie par*

$$\tilde{S}_2(r) = \frac{1}{r} \log \frac{p(\mathcal{X}_1^r|\hat{\theta}_1)}{p(\mathcal{X}_1^r|\hat{\theta}_2)} + \frac{1}{n-r} \log \frac{p(\mathcal{X}_{r+1}^n|\hat{\theta}_2)}{p(\mathcal{X}_{r+1}^n|\hat{\theta}_1)}$$

Dans le cas de modèles Gaussiens, $\tilde{S}_2(r)$ est égale à la divergence de Kullback entre les distributions de paramètres $\hat{\theta}_1 = (\hat{\mu}_1, \hat{\Sigma}_1)$ et $\hat{\theta}_2 = (\hat{\mu}_2, \hat{\Sigma}_2)$, donnée par:
$\frac{1}{2}\Big[Tr(\hat{\Sigma}_1\hat{\Sigma}_2^{-1} + \hat{\Sigma}_2\hat{\Sigma}_1^{-1}) - 2d +$
$(\hat{\mu}_1 - \hat{\mu}_2)^T(\hat{\Sigma}_1^{-1} + \hat{\Sigma}_2^{-1})(\hat{\mu}_1 - \hat{\mu}_2)\Big]$

Les indices S_1 et S_2 sont respectivement obtenus à partir des statistiques $\Lambda_1(n)$ et $\Lambda_2(n)$, en faisant abstraction du supremum ($\sup_{r=m,\cdots,n-m}$) et des fonctions de pondération (ψ_1 et ψ_2), avec les correspondances $n = 2L$ et $t = r = L$. Ensuite, on opère une normalisation par L, dans l'idée d'une mesure d'amplitude de rupture qui, en moyenne, est plus robuste à la taille de la fenêtre d'analyse.

5 Expériences

Dans nos expériences, le problème de la segmentation d'une bande sonore en plages de parole versus plages de musique, a été considéré. La base de données est constituée de morceaux de signaux de parole et de musique d'environ 4.8 secondes chacun, enregistrés sur différentes stations (une vingtaine environ, émettant à Rennes), échantillonnés sur 16 bits à 16 kHz.
Chaque morceau de signal est ensuite découpé en trames de 16 ms, paramétrisés chacune par un vecteur de 25 Composantes Principales du Spectre (CPS). Dans un premier temps, nous avons généré un signal de dimension 25, obtenu par concaténation de segments simulés. Chaque segment simulé correspond à environ 300 réalisations indépendantes d'une variable aléatoire de loi Gaussienne à covariance pleine. Pour chaque segment, un modèle différent est estimé sur un ensemble de vecteurs de 25 CPS extraits d'un morceau de signal de parole ou de musique. Le signal testé est obtenu en alternant les segments générés selon les modèles de parole et ceux générés selon les modèles de musique. Le signal ainsi simulé compte environ 200 ruptures. Un second signal, obtenu en concaténant directement les vecteurs de CPS extraits des morceaux de parole et de musique, a également été généré.
La méthode de segmentation décrite à la section 3, est appliquée au signal des segments simulés et au signal des segments réels, avec des modèles de segments à une Gaussienne, en utilisant les deux indices de rupture et pour différentes valeurs de la taille de la fenêtre d'analyse. Les performances d'une segmentation sont mesurées par le pourcentage de ruptures non détectées et le pourcentage de fausses alarmes. Nous tolérons une erreur de 5 trames sur la localisation de l'instant de rupture. C'est à dire que chaque instant de rupture $r^\star$ est redéfini par $\tilde{r} = \arg\max_{t=r^\star-5,\cdots,r^\star+5} C(t)$, qui devient alors l'instant de rupture de référence. La figure 1 représente les performances obtenues sur les deux signaux.

6 Interprétation des résultats

On observe sur la figure 1, qu'indépendamment de l'indice de rupture utilisé, les performances sont généralement

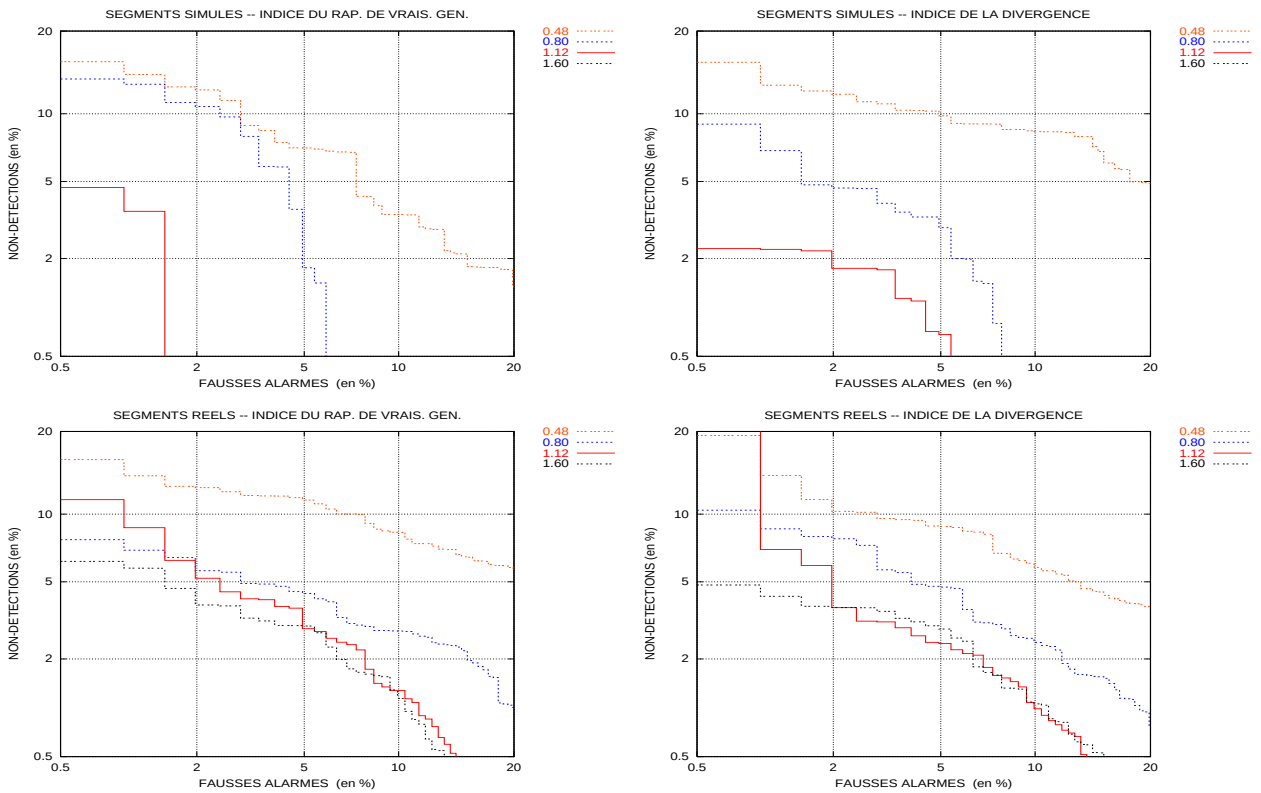

FIG. 1: *Performances pour differentes durées de la fenêtre d'analyse. En haut, les performances obtenues sur le signal des segments simulés en utilisant l'indice du RVG (à gauche) et en utilisant l'indice de la divergence (à droite). Les taux d'erreurs sont nuls quand la durée de la fenêtre d'analyse est égale à* 1.60 *s. En bas, les performances obtenues sur le signal des segments réels en utilisant l'indice du RVG (à gauche) et en utilisant l'indice de la divergence (à droite).*

croissantes en fonction de la durée de la fenêtre d'analyse. De meilleures performances sont observées sur le signal à segments simulés. Pour le signal à segments simulés, les taux d'erreurs deviennent nuls à partir d'une certaine durée de la fenêtre d'analyse (1, 60 s), alors que pour le signal à segments réels les performances on tendance à se stabiliser. Cette différence de performances peut se justifier d'une part, par la cohérence entre le type de distributions (Gaussien) des segments et la famille de distributions utilisée par le procédé de segmentation, dans le cas du signal des segments simulés. D'autre part, parce que sur le signal des segments réels, en plus des ruptures de type parole/musique, intervient la non-stationnarité des segments générant d'autres ruptures. Néanmoins, les performances sur le signal des segments réels sont satisfaisantes, avec un «*Equal Error Rate*» (EER) inférieur à 5% dès que la durée de la fenêtre d'analyse excède 0, 80 s.
Nous n'observons pas de différence significative entre les deux indices de ruptures. Cependant, on note une plus grande tendance à la non-détection pour l'indice du RVG, tandis que l'indice de la divergence s'avère plus prédisposé aux fausses alarmes.

7 Conclusion

Un méthode de segmentation de signaux sonores, par détection de ruptures a été présentée et évaluée sur un signal obtenu par concaténation de réels segments de parole et de musique en alternance. Ces expériences ne montrent pas une différence significative en terme de performances, entre deux statistiques de test d'existence d'une rupture: le rapport de vraisemblance généralisé et une seconde statistique présentée ici, celle de la divergence. Des performances assez satisfaisantes ont été obtenues en utilisant une famille de modèles assez simple, les modèles Gaussiens. Dans nos futurs travaux, cette méthode sera évaluée sur des signaux présentant des transitions plus naturelles entre les différents segments.

Références

[Bra83] Brandt (A. Von). – Detecting and estimating parameter jumps using ladder algorithms and likelihood ratio tests. *In: ICASSP*, pp. 1017–1020. – 1983.

[DAOS88] Delyon (Bernard), Andre-Obrecht (Régine) et Su (H. Y.). – Expériences en vue du décodage acoustico phonétique à partir d'une recherche d'événements acoustiques et d'un codage vectoriel. *Journal of Acoustique*, vol. 1, septembre 1988, pp. 229–201.

[DP86] Deshayes (Jean) et Picard (Dominique). – Off-line statistical analysis of change-point models using non parametric and likelihood methods. *In: Detection of Abrupt Changes in Signals and Dynamical Systems*, éd. par Basseville (M.) et eds (A.Benveniste), pp. 103–168. – Springer-Verlag, 1986.

[SBD98] Seck (Mouhamadou), Bimbot (Frédéric) et Delyon (Bernard). – Classification parole/musique de signaux sonores paramétrisés par les composantes principales du spectre. *In: Journées d'Études sur la Parole (JEP)*, pp. 331–334. – juin 1998.