Segmentation multibande adaptée basée sur le Critère Entropique Local pour le codage audio

Gilles GONON[1], Silvio MONTRÉSOR[2], Marc BAUDRY[1]

[1]Laboratoire d'Informatique de l'Université du Maine
Av. F. Laennec, 72085 Le Mans Cedex

[2]Laboratoire d'Acoustique de l'Université du Maine, UMR CNRS 6613,
Université du Maine, 72085 Le Mans Cedex
Gilles.Gonon@univ-lemans.fr, Silvio.Montresor@univ-lemans.fr, Marc.Baudry@univ-lemans.fr

Résumé – Ce travail présente une nouvelle approche pour la segmentation des signaux audios. Le détecteur utilisé est non paramétrique et basé sur le Critère Entropique Local appliqué aux sous-bandes issues d'une analyse multirésolution, la transformée en ondelettes discrète (TOD). L'utilisation de la TOD permet d'augmenter la diversité des ruptures détectées et le taux de bonnes détections. Un post-traitement permettant de réduire le nombre de fausses alarmes est aussi présenté. Les résultats sont appliqués à un signal de simulation multicomposante bruité.

Abstract – This paper presents a new method for the segmentation of audio signals. We used a non parametric detector based on the Local Entropic Criterion applied to the subbands resulting from a Discrete Wavelet Transform. The use of such a multiresolution analysis increases the diversity of changes detected and so the right detections rate. We also present a post-processing of the criterions that reduces the number of false alarms detected. Results are discussed on a multicomponants simulation signal with noise.

1 Introduction

Dans le cadre du codage audio, une segmentation temporelle adaptée au signal permet d'améliorer la qualité des signaux codés en réduisant des effets tels que le pré-écho [1]. Les normes MPEG-2 et AC-3 utilisent une transformation du signal adaptée temporellement mais n'imposent pas de méthodes de segmentation. Nous proposons ici une méthode de segmentation efficace des signaux audios basée sur une approche non paramétrique.

Les signaux audios couvrent une vaste gamme de fréquences, ce qui conduit à des comportements très divers au niveaux des instationnarités. Aux deux extrêmes il y a d'une part les transitions de type *percussives* qui correspondent à d'importantes variations d'énergie à court terme et d'autre part les transitions de type *legato* pour lesquelles l'énergie reste constante mais dont le spectre varie aléatoirement. Entre les deux toutes les combinaisons sont possibles (crescendo, glissando), et le signal est presque toujours multicomposante.

Devant la diversité de ces instationnarités une lacune des détecteurs de ruptures paramétriques est de considérer un modèle de ruptures [2] ne permettant pas de s'adapter aux différents contextes des ruptures. Une alternative proposée dans [3] permet de détecter les changements abruptes dans les signaux musicaux mais tous les types de ruptures ne sont pas détectés, notamment lorsque le signal est multicomposante.

Afin de pallier ce problème, la méthode proposée ici consiste à faire premièrement une analyse multirésolution puis à détecter les ruptures présentes dans chaque sous-bande à l'aide du Critère Entropique Local, adapté à chaque résolution. Un post-traitement permet ensuite une localisation précise des ruptures détectées aux différentes résolutions afin d'obtenir une segmentation automatique des signaux. Les résultats des différentes segmentations comparés sur un signal de simulation multicomposante montrent le gain en terme de bonnes détections d'apparition d'évènements.

2 Le Critère Entropique Local

2.1 Définition

Le Critère Entropique Local(CEL), défini dans [4], est une fonction temporelle mesurant les variations de la concentration d'énergie du spectre à court terme du signal. La mesure de concentration de l'énergie utilisée est l'entropie de Shannon appliquée à la transformée de Fourier discrète du signal.

En notant $X(k)$ la transformée de Fourier discrète normalisée d'un signal $x(n), (k,n) \in [0, N-1]$

$$X(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) W_N^{nk} \quad \text{, où } W_N^{nk} = e^{-2j\pi \frac{nk}{N}}. \quad (1)$$

L'entropie de Shannon de x sur l'intervalle $[0, N-1]$ notée $E_{x[0,N-1]}$, est définie par

$$E_{x[0,N-1]} = -\sum_{k=0}^{N-1} |X(k)|^2 \log |X(k)|^2. \quad (2)$$

Le CEL est alors défini comme une différence relative d'entropie entre une fenêtre glissante de longueur N et ses 2

sous-fenêtres de longueur $\frac{N}{2}$. En notant respectivement E_{xc}, E_{xg} et E_{xd} les entropies de la fenêtre principale et de ses deux sous-fenêtres gauche et droite, définies par (Fig. 1)

$$\begin{aligned} E_{xc}(n) &= E_{x[n-\frac{N}{2},n+\frac{N}{2}-1]}, \\ E_{xg}(n) &= E_{x[n-\frac{N}{2},n-1]}, \\ E_{xd}(n) &= E_{x[n,n+\frac{N}{2}-1]}, \end{aligned} \tag{3}$$

le CEL est défini pour un signal de longueur M sur l'in-

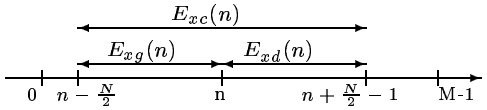

FIG. 1: *Notations des entropies pour le calcul du CEL*

tervalle $[\frac{N}{2}, M-\frac{N}{2}-1]$ par la formule suivante:

$$CEL_x(n) = \frac{E_{xc}(n) - (E_{xg}(n) + E_{xd}(n))}{|E_{xc}(n)|}. \tag{4}$$

La taille de la fenêtre définit le *contexte* d'analyse. Afin de réduire l'écart type de l'entropie en fonction de la fréquence, chaque tranche est apodisée par une fenêtre de Hamming. Ceci permet de supprimer les différences de spectre causées par le calage de la fréquence sur une harmonique de la TFD qui modifie la valeur de l'entropie résultant de la somme de tous les coefficients.

Le terme de normalisation du dénominateur du CEL est choisi de telle sorte que le CEL ait une valeur proche de -1 pour un signal stationnaire monocomposante. Pour un signal nul, le CEL conduit à une forme indéterminée du type $\frac{0}{0}$; aussi nous fixons la valeur du CEL à -1 pour un signal nul.

2.2 Propriétés

Le CEL est construit de manière similaire, au facteur de normalisation prêt, aux tests entropiques utilisés dans les algorithmes de recherche de meilleure base pour la décomposition en paquets d'ondelettes. Deux principales propriétés du CEL basées sur les mêmes principes sont utilisées comme indices pour effectuer la segmentation automatique:

Une zone stationnaire ou *stable* est caractérisée par un CEL < 0 et proche de la valeur -1. (conservation du père pour la meilleure base). Ceci est dû au fait que l'énergie est deux fois plus concentrée sur une fenêtre de longueur N que sur les fenêtres de longueur $\frac{N}{2}$, d'où en raison de la concavité de la fonction entropie

$$E_{xc}(n) < (E_{xg}(n) + E_{xd}(n))$$

Une zone *instable* est caractérisée par un CEL > 0. Une perturbation de spectre traduisant l'apparition d'une rupture dans la demi-fenêtre de droite va alors justifier entropiquement la séparation en deux fenêtres au sens où

$$E_{xc}(n) > (E_{xg}(n) + E_{xd}(n))$$

Ceci est dû au fait que l'entropie de Shannon est maximisée pour un signal équiprobable. Ici le spectre étant assimilé à la densité de probabilité des fréquences, la propriété d'additivité de la transformée de Fourier induit que la fenêtre N est constituée de la somme des deux spectres des demi fenêtres et n'est plus deux fois plus concentrée comme dans le cas d'un signal stationnaire. Le spectre de la fenêtre devient ainsi plus équiprobable.

En pratique, lorsque les signaux sont bruités par exemple, le CEL ne prend que rarement des valeurs positives et une zone instable est alors délimité par un maximum local de la courbe, situé au dessus d'un seuil préalablement fixé.

2.3 Conditionnement et contexte d'analyse

Le CEL fournit une segmentation efficace des signaux lorsque le contexte d'analyse est connu, comme c'est le cas par exemple de la parole en bande téléphonique [4]. Cependant, d'après la construction du détecteur, les résultats obtenus sont fortement liés au contexte d'analyse. La diversité dans la nature des ruptures et leur fréquence d'apparition influent sur le conditionnement du problème au sens des variations du CEL et du diagnostique pour la segmentation. C'est-à-dire que pour un contexte donné, il n'est pas suffisant de fixer les points de segmentation au niveau des maximas du CEL supérieurs à un seuil fixé. Différents comportement existent suivant les types de ruptures concernées, et le problème devient mal conditionné lorsque qu'un comportement est *masqué* par un autre plus évident. Parmi les différents comportements, on distingue aux deux extrêmes:

- les transitions entre deux zones stationnaires. Le CEL indique dans ce cas là un maximum localisé précisément au niveau de la transition, les deux demi fenêtres étant chacunes stationnaires tandis que la fenêtre centrale est la somme des ces deux demi spectres. La détection est robuste au bruit mais le seuil de segmentation doit varier en fonction du RSB (Fig. 2).
- les ruptures de type pic de Dirac ou à support temporel petit devant le contexte d'analyse. La variation de CEL est moins importante que dans le cas précédent et l'incertitude sur la localisation et égale à la moitié du contexte, car le module de la transformée de Fourier utilisé pour la mesure entropique ne contient pas d'information temporelle sur la position de la rupture. L'ajout de bruit sur un pic de dirac entraîne la disparition de la détection (Fig. 3).

En présence de bruit, il est nécessaire de modifier le seuil de détection (typiquement entre 0.5 et 0.6) pour conserver le taux de bonnes détections.Il est à noter qu'un bruit blanc gaussien centré a un CEL de distribution gaussienne centrée en

$$-1 + \frac{2\log 2}{\log N}, \text{où } N\text{est le contexte.}$$

et de variance inversement proportionnelle à N. La modification du seuil est liée au terme $\frac{2\log 2}{\log N}$ correspondant à la moyenne du CEL d'un bruit qui diminue les maximas locaux du CEL.

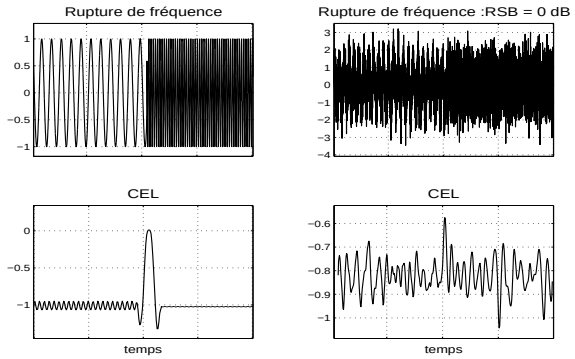

FIG. 2: Bas: Courbes du CEL obtenues respectivement pour un signal de type échelon de fréquence, sans bruit (haut gauche) et pour un RSB de 0dB (haut droit)

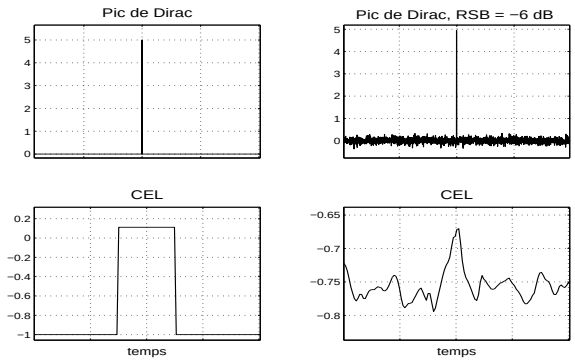

FIG. 3: Bas: Courbes du CEL obtenues respectivement pour un signal de type pic de dirac, sans bruit (haut gauche) et pour un RSB de -6dB (haut droit)

Le compromis effectué sur le contexte d'analyse afin d'avoir une bonne résolution fréquentielle ainsi qu'une bonne localisation temporelle des ruptures ne permet pas de s'adapter aux différentes variations présentes dans les signaux audios. L'utilisation d'une analyse multirésolution ajoute un degré de liberté sur le contexte en permettant une adaptation différente pour chaque résolution.

2.4 CEL en analyse multirésolution

La Transformée en Ondelette Discrète dyadique (TOD) fournit une segmentation fréquentielle du signal en séparant les fluctuations rapides et lentes d'un signal [5]. Elle est équivalente à une analyse par un banc de filtre en octave (Fig. 4), les réponses impulsionnelles des différents filtres correspondant aux ondelettes mères de chaque résolution. La précision sur la localisation étant proportionnelle à la taille des ondelettes mères de chaque résolution, elle diminue d'un facteur 2 du fait que le passage à la résolution inférieure se fait par dilatation d'un facteur 2 de l'ondelette mère.

L'application du CEL à chaque résolution de la TOD apporte deux améliorations. D'une part le problème de détection de changement sur le spectre à court terme devient mieux conditionné car l'entropie de chaque résolution, ou sous bande, n'est plus maximisée par $\log N$ mais par $\log \frac{N}{2^{(D-d)}}$, où D est la profondeur de la TOD et d la

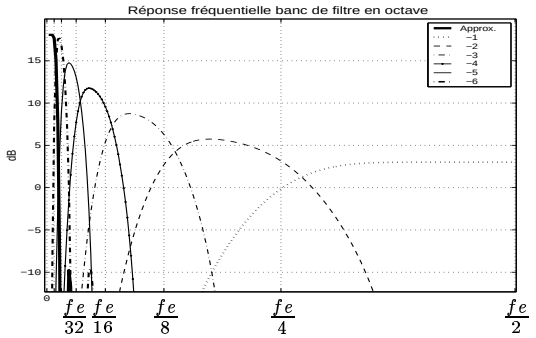

FIG. 4: Réponse fréquentielle du banc de filtre en octave équivalent à la TOD de profondeur 6.

résolution. Les variations au sein d'une sous bande seront donc moins noyées dans le bruit. D'autre part, la possibilité d'adapter le contexte du CEL aux différentes résolutions de la TOD permet d'obtenir une localisation temporelle précise des évènements rapides dans les hautes fréquences, mais aussi une bonne résolution fréquentielle dans les moyennes et basses fréquences. Les ruptures de type pic de dirac sont transformées par l'ondelette analysante ce qui modifie la réaction du CEL et facilitant leur localisation, au sens ou un maximum apparaît. Le choix du contexte est présenté à la section 4 pour les différents signaux analysés.

Un post traitement est alors nécessaire pour constituer la segmentation finale du signal en regroupant les différentes ruptures détectées à chaque résolution.

3 Diagnostique de segmentation

Le CEL fournissant un indice de désordre dans le signal, la segmentation repose entièrement sur le diagnostique des ruptures. De plus le terme de segmentation est souvent lié au contexte de l'application. Dans une application comme le codage audio, on cherche à à isoler des zones de comportement différents et non à détecter des points précis de rupture. Dans une application telle que l'extraction de parole dans le bruit, on cherchera des points précis de début et de fin de parole. C'est l'application qui fixe les impératifs du diagnostique.

La détection des ruptures est effectuée en deux étapes constituant le post-traitement, à partir des courbes CEL obetnues pour chaque sous bandes.

1. Au sein de chaque courbe CEL, un seuil inférieur à 0 (typiquement -0.3) est fixé. Pour chaque zone du CEL supérieure au seuil tous les maximas locaux sont conservés. Une incertitude sur la localisation subsiste parfois (comme pour la détection des diracs voir figure 3) lorsque deux maximas sont détectés dans un intervalle inférieur au contexte. Afin de lever cette incertitude, la plus grande valeur du CEL est considérée comme maximum le plus vraisemblable au sein d'un intervalle de la taille du contexte.

2. Entre les courbes CEL des différentes résolutions, l'incertitude sur la localisation dépend maintenant de la taille de l'ondelette analysante. L'incertitude est donc fixé par la longueur de l'ondelette père (filtre QMF) à la résolution 0 et elle est dilatée d'un facteur 2 à chaque résolution

inférieure. Entre deux résolutions, lorsque deux ruptures sont à l'intérieur de l'intervalle d'incertitude, la rupture de résolution la plus fine est conservée car c'est la mieux localisée.

4 Résultats

Pour le codage audio, les signaux sont souvent peu bruités mais très complexes spectralement. Le signal de simulation analysé est constitué de 4 sinusoïdes de supports temporels différents et de 2 pics de diracs (Fig. 6). Les courbes CEL obtenues pour chaque résolution de la TOD sont données, avec les coefficients d'ondelettes correspondants à la figure 5. Aux résolutions les plus basses, l'incertitude sur la localisation devient supérieure à la taille de la sous-bande, aussi les résultats sur ces sous bandes ne sont pas utilisées pour le post-traitement.

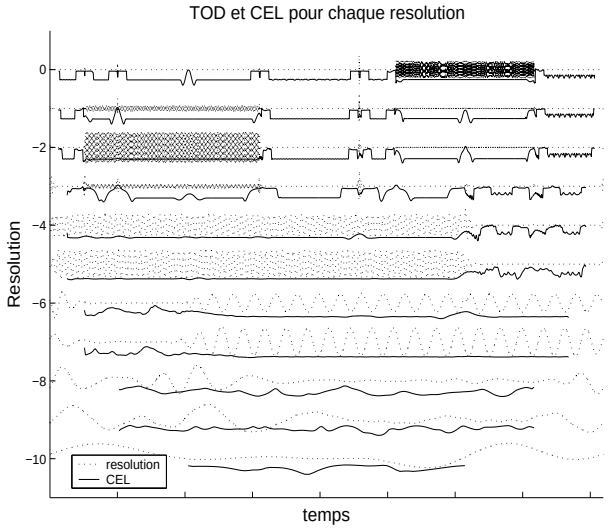

FIG. 5: Différentes résolutions de l'analyse par TOD (pointillé) et CEL correspondant (continu).

Les segmentations obtenues par les 3 méthodes du CEL, sans TOD, avec TOD sans post-traitement et avec TOD et post-traitement sont comparées à la figure 6. La comparaison montre l'apport de la multirésolution sur le taux de bonne détection, puisque sur l'exemple proposé les occurrences des sinus ainsi que les diracs sont détectés, tout en permettant d'augmenter le seuil de détection. Cependant le taux de fausses alarmes augmente aussi et il est nécessaire d'appliquer le post traitement pour le réduire.

Pour un RSB de 5dB, le taux de bonnes détection reste satisfaisant avec un seuil de -0.6. Le bruit influence principalement le taux de fausses alarmes et la localisation des pics de Diracs, car les maximums du CEL peuvent être perturbés localement sur les paliers de détection d'un dirac.

5 Conclusion

Le détecteur proposé basé sur le critère entropique local utilisé en analyse multirésolution permet d'obtenir une

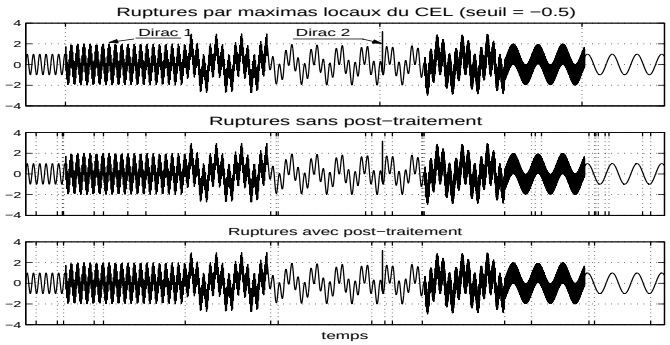

FIG. 6: Comparaison des segmentations obtenues sur le signal de simulation (4 sinus + 2 diracs) par : CEL (haut), CEL multirésolution sans post-traitement (centre), CEL multirésolution et post traitement (bas).

segmentation du signal fiable pour différents types de ruptures existant dans les signaux audios. Ceci est dû à la possibilité d'adaptation du contexte pour différentes résolutions et au fait que le CEL est un détecteur efficace quand le contexte est connu.

Les premières analyses montrent que la prise en compte de plusieurs niveaux de résolutions améliore sensiblement le taux de bonnes détection. Il est cependant nécessaire de regrouper les ruptures se retrouvant à différentes résolutions avec des localisations différentes par un post traitement prennant en compte l'erreur de localisation temporelle due à la TOD.

Les résultats obtenus sur un signal multicomposante montrent que les différentes ruptures sont détectées et que le post traitement réduit le taux de fausses alarmes. Une étude formelle des propriétés statistique du CEL reste cependant nécessaire pour optimiser le diagnostique de la segmentation.

Références

[1] Seymour Shlien. The modulated lapped transform, its time-varying forms, and its applications to audio conding standards. *IEEE trans. on speech, audio processing*, 5(4):359–366, July 1997.

[2] GDR TdSI GT2. Reconnaissance et ruptures. segmentation de signaux : Fiches descriptives d'algorithmes, 1991.

[3] H. Laurent, E. Hitti, and M.F. Lucas. Abrupt changes detection in the time-scale and in the time-frequency planes : a comparative study. In *TFTS, IEEE-SP International Symposium on Time-frequency and Time-scale Analysis*, pages 581–584, 1998.

[4] Imad Abdallah, Silvio Montresor, and Marc Baudry. Un algorithme récursif pour la segmentation des signaux de parole basé sur un critère entropique local. In *4ème Congrès de la Société Française d'Acoustique (SFA)*, pages 85–88, Avril 1997.

[5] Albert Cohen. *Ondelettes et traitement numerique du signal.* Masson, 1992.