Détection de ruptures à l'aide des « SVM 1 classe » pour la segmentation des signaux sonores musicaux

Stéphane Rossignol, Manuel Davy

LAGIS/CNRS/INRIA-FUTURS « SequeL »,
Cité Scientifique,
BP 48, 59651 Villeneuve d'Ascq Cedex, France
Stephane.Rossignol@ec-lille.fr,Manuel.Davy@ec-lille.fr

Résumé – Ce travail s'inscrit dans le cadre général de la segmentation des sons en notes ou en phones (suivant la nature du son : instrumental, voix chantée, ou parole), ou, plus globalement, de leur segmentation en « parties quasi stationnaires ». Nous nous plaçons dans le cas où, à long terme, aucune hypothèse sur la nature du signal n'est faite : le son peut être de la musique tonale monophonique, mais il peut aussi être de la parole ou de la musique polyphonique ; il peut être harmonique ou inharmonique (extrait de castagnettes) ; les notes courtes (moins de 100 millisecondes) et le vibrato doivent être pris en compte ; etc. Le travail présenté dans [2] et [6] était un premier pas vers ce but. Le système d'extraction de descripteurs audio et de segmentation présentés dans cet article constitue un pas supplémentaire en direction de notre objectif. Il a aussi pour but d'améliorer la robustesse du système proposé dans [2] et [6].

Abstract – We deal with segmentation into *note* and/or into *phone* (according to the nature of the sound: instrumental part or singing voice excerpt or speech) or more generally into "stable" parts. There should have no restriction (see [9]) on the signal to be segmented: the sound can be monophonic music, but it can also be speech or polyphonic music; it can be harmonic or inharmonic (castanets or drums for example); fast notes (less than 100 milliseconds duration for example) and vibrato should be taken into account. The system presented in [2] and [6] was a *work in progress* towards this goal. The segmentation and feature extraction system described here is another step towards the more general system. It aims also to improve the robustness of the system described in [2] and [6].

1 Introduction

L'indexation et la segmentation des sons, quels qu'ils soient (parole, musique, bruit...), d'où qu'ils viennent (radios, bandes son de films, CD...), sont des domaines qui sont en plein essort, notamment du fait de la définition de standards comme MPEG-7. Il s'agit de définir un ensemble d'outils de description de « contenus » multimédia, pour en faciliter la recherche, l'identification et aussi la manipulation. Dans cet article, nous proposons et nous étudions des techniques pour segmenter et, dans une moindre mesure, étiqueter les signaux sonores. Nous montrons notamment l'efficacité des méthodes à noyaux (plus spécifiquement, les « SVM 1 classe ») pour ce faire.

La segmentation est effectuée en trois étapes. Premièrement, des descripteurs audio sont extraits du signal audio à partir de trames bien localisées en temps (section 2). Deuxièmement, les ruptures sont détectées dans l'espace des descripteurs audio grâce à l'algorithme *Kernel Change Detection* (KCD) [3] (section 3). Troisièmement, afin que d'obtenir les marques de segmentation, la sortie de l'algorithme KCD est seuillée automatiquement (section 4). Puis, dans la section 5, les corpus de test sont présentés. Dans la section 6, tout d'abord des résultats sur un petit ensemble de test sont donnés. Des résultats plus complets, sur un plus grand ensemble, sont donnés ensuite. Ces tests permettent de comparer les performances du système actuel avec celles du système décrit dans [2] et [6]. Finalement, la section 7 conclut cet article.

2 Les descripteurs audio considérés

La segmentation des signaux audio se fait essentiellement sur la base de caractéristiques fréquentielles et perceptives du son. Celles-ci peuvent, dans une certaine mesure, être quantifiées sous la forme de descripteurs audio, extraits pour la trame courante. Une trame est un segment de signal modulé par une fenêtre de type Hamming, par exemple, de durée 50 ms ici. Dans les résultats présentés ci-dessous, l'écart entre deux trames successives est de 10 ms. Six descripteurs audio sont retenus ici. On pourra se reporter à [4] pour une description complète.

- La **fréquence fondamentale** f_0 est estimée selon la technique developpée dans [5]. Les résultats présentés dans [2] avaient été obtenus avec l'estimateur de f_0 *additive*, développé à l'IRCAM (voir [8]).
- L'**énergie**, calculée dans le domaine temporel.
- Le **barycentre spectral**, qui est défini comme le centre de gravité du spectre d'amplitude de chaque trame.
- La **largeur de bande,** qui est définie de façon similaire à un écart-type, lorsque le spectre d'amplitude est vu comme une densité de probabilité.
- Les **énergies par bandes,** qui sont définies comme

l'énergie du signal, pour une trame donnée, dans plusieurs bandes de fréquences. Les trois mêmes bandes que dans [4] sont considérées.

- Les n **coefficients cepstraux MFCC** (*Mel Filter Cepstrum Coefficients*). Les MFCC sont très communément utilisés en reconnaissance automatique de la parole, et sont basés sur une modélisation de l'audition humaine.

Lorsque ces descripteurs audio ont été extraits, ils sont rassemblés dans un vecteur de dimension $7 + n$, noté x_t où t est l'instant de localisation de la trame.

3 L'algorithme *Kernel Change Detection* (KCD)

La détection des ruptures est effectuée séquentiellement dans l'espace des descripteurs audio. Pour un instant d'analyse t, on considère l'ensemble des m descripteurs audio passés et futurs, soit respectivement :

$$X_{1,t} = \{x_{t-m}, \ldots, x_{t-1}\} \text{ et } X_{2,t} = \{x_{t+1}, \ldots, x_{t+m}\}.$$

Typiquement, m est pris de l'ordre de 30, ce qui correspond à des portions de 0,3 seconde. L'espace des descripteurs audio est noté $\mathcal{X}$ et l'on a $\mathcal{X} = \mathbb{R}^{n+7}$.

À l'instant t, pour chacun des deux ensembles $X_{1,t}$ et $X_{2,t}$, un estimateur de support de densité SVM (dit « SVM une classe » – 1-SVM) est appris. Nous ne considérons pas ici l'algorithme classique des « SVM deux classes », mais nous faisons l'hypothèse que les m données d'apprentissage appartiennent à la même classe, avec l'étiquette $+1$. Le 1-SVM estime la région $\mathcal{R}$ dans $\mathcal{X}$ de volume minimum contenant au moins $(1-\nu)m$ données, ν déterminant asymptotiquement la proportion de données hors volume. La frontière de $\mathcal{R}$ est recherchée dans un espace de Hilbert à noyau reproduisant (RKHS) de noyau $k(\cdot,\cdot)$ où le problème de minimisation de volume peut être reformulé sous la forme de minimisation du volume d'une calotte sphérique en dimension m (voir [3]). Il s'agit d'un problème d'optimisation qui admet la forme duale suivante :

$$\text{minimiser} \quad \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j k(x_i, x_j) \tag{1}$$

$$\text{par rapport à} \quad \alpha_1, \ldots, \alpha_m \tag{2}$$

$$\text{avec} \quad \sum_{i=1}^{m}\alpha_i = 1 \quad \text{et} \quad \forall_i : 0 \leq \alpha_i \leq \frac{1}{\nu m} \tag{3}$$

où les couples (x_i, α_i) déterminent la région $\mathcal{R}$ dans $\mathcal{X}$.

Grâce à cette technique, il est possible d'apprendre à chaque instant t les régions $\mathcal{R}_{1,t}$ et $\mathcal{R}_{2,t}$. Ces deux régions peuvent être comparées à l'aide de la mesure introduite dans [3]. En d'autres termes, une mesure de dissimilarité $I(t)$ est calculée à partir des couples (x_i, α_i) pour $X_{1,t}$ et $X_{2,t}$ (son calcul n'est pas détaillé ici faute de place). Cet « index de non-stationarité » $I(t)$ est utilisé pour détecter les ruptures par seuillage.

4 Seuillage

La segmentation d'un signal est efficace si l'indice $I(t)$ construit a de « bonnes » caractéristiques. En particulier, il faut que son évolution temporelle présente des pics hauts et peu larges quand une transition – ou rupture – entre deux notes a lieu, et que sa moyenne et sa variance restent faibles en l'absence de rupture. Cet index doit être seuillé automatiquement, c'est-à-dire que les maximums locaux les plus significatifs doivent être sélectionnés. Dans la littérature, de nombreuses techniques de seuillage automatique ont été proposées (voir [7]), en particulier pour le traitement des images. Dans notre cas, les pics dus aux transitions sont rares et leur variance est grande.

Dans un premier temps, la règle des 3σ a été retenue. Pour cela, nous supposons que le signal $I(t)$ est Gaussien en dehors des pics, et nous déterminons sa moyenne μ et son écart-type σ sur la base des $\eta = 90\%$ plus petites valeurs de $I(t)$. Un pic est détecté si $I(t) \geq \mu + 3\sigma$. La méthode est relativement robuste suivant η (voir [2], annexe B).

Dans un second temps, la valeur du seuil est validée en examinant la courbe COR (qui est parcourue en faisant varier le seuil : voir la figure 1).

5 Données

5.1 Premier ensemble de sons

Nous présentons ici les résultats de segmentation pour les quatre enregistrements utilisés dans [6]. Ils comprennent :

- Un extrait de flûte, « flute.wav », enregistré en chambre anéchoïque. Il comporte très peu de réverbération : considérant une transition entre deux notes donnée, elle ne s'étale pas sur la note qui la suit. De plus, ce son est presque parfaitement harmonique et le taux de modulation est faible.
- Un extrait de clarinette, « brahms2.wav ». Il comprend des notes très courtes (moins de 100 millisecondes).
- Un extrait de violon, « Violon2.wav », très bruité du fait d'un enregistrement à bas niveau. De plus, on peut entendre le bruit des pages de la partition quand elles sont tournées.
- Un extrait de voix chantée, « voiceP.wav ». L'amplitude du vibrato et celle du trémolo sont très importantes. Il faut de plus noter que la même voyelle est chantée tout le long de l'extrait.

5.2 Second ensemble de sons

En plus des 4 sons ci-dessus, 14 sons du CD Sqam (« Sound Quality Assessment Material », édité par : « European Broadcasting Union ») et un son de piano sont considérés. Les sons du CD Sqam suivants sont pris en compte :

- **3.wav** : extrait d'un gong électronique artificiel. Ce son est formé de douze fois la même note. Chaque note est composée d'une seule sinusoïde, amortie exponentiellement avec le temps. La fréquence de cette

sinusoïde est $f_0 = 100\ Hz$. Les notes sont rassemblées en quatre groupes de trois notes très rapprochées. Les groupes sont séparés par des silences de durée 1 s. Pour chaque groupe, nous avons 4 transitions à détecter. La durée de chaque note est $1,3\ s$.

- **6.wav** : extrait d'un autre gong électronique artificiel. Ce son est formé de huit fois la même note. Chaque note est composée d'une seule sinusoïde, amortie exponentiellement avec le temps. La fréquence de cette sinusoïde est $f_0 = 475\ Hz$. Un vibrato est présent. L'amplitude du vibrato est $20\ Hz$. La durée de chaque note est $1,3\ s$. Un silence de $0,5\ s$ sépare deux notes successives. Ainsi, nous avons 16 transitions à détecter.
- **8a.wav** : extrait d'un violon. Un vibrato, très petit, est présent. Les notes jouées vont du sol_2 ($f_0 = 196\ Hz$) au sol_5 ($f_0 = 1568\ Hz$).
- **11a.wav** : extrait d'une contrebasse. Les notes jouées sont très graves : les fréquences fondamentales sont comprises entre $61,7\ Hz$ (si_0) et $392\ Hz$ (sol_3).
- **14a.wav** : extrait d'un hautbois. Les notes jouées sont plutôt aiguës. Les fréquences fondamentales sont comprises entre $293,66\ Hz$ (ré_3) et $1174,64\ Hz$ (ré_5).
- **16a.wav** : extrait d'une clarinette. Les notes sont longues.
- **17a.wav** : extrait d'une clarinette basse. Les notes sont longues.
- **19a.wav** : extrait d'un contre-basson. Cet extrait est composé de notes très graves : entre $32,7\ Hz$ (do_0) et $130,8\ Hz$ (do_2). La taille t_{SIG} des fenêtres d'analyse doit être choisie plus grande que pour les autres sons : pour le do_0, t_{SIG} doit être de l'ordre de 80 ou 100 millisecondes.
- **20a.wav** : extrait d'un saxophone. Un léger vibrato est présent sur la dernière note.
- **20b.wav** : autre extrait d'un saxophone. Pour les deux extraits de saxophone, la chute de la dernière note est très lente, il est donc difficile de déterminer à quel moment elle finit. Les résultats donnés par l'estimateur de f_0 deviennent de plus en plus chahutés.
- **29.wav** : extrait d'une grosse caisse. Il est composé de six coups séparés par environ 3 secondes.
- **30.wav** : extrait d'une timbale. Il est composé de dix coups séparés par au moins 1,5 seconde. Nous entendons une hauteur, mais aussi des battements : certaines sinusoïdes ont donc des fréquences très proches. Il y a des partiels perturbateurs. L'estimateur de f_0 ne parvient pas à nous donner une fréquence fondamentale.
- **39a.wav** : extrait d'un piano. Il est composé de dix notes. La chute de la dernière note est très lente.
- **42a.wav** : extrait d'un accordéon. Il est composé de vingt-quatre notes, dont certaines sont très courtes (moins de 100 millisecondes).
- **43a.wav** : extrait d'un orgue. Il est composé de sept notes.

L'extrait de piano est le suivant :

- **piano2.wav** : une note de piano. Le bruit de l'étouffoir quand il se referme sur la corde à la fin de la note est audible.

Pour les sons **8a.wav**, **11a.wav**, **14a.wav**, **16a.wav**, **17a.wav**, **19a.wav**, **20a.wav**, **39a.wav** et **43a.wav**, un arpège ascendant est joué ; alors que pour les sons **20b.wav** et **42a.wav**, une mélodie est jouée. Tous les sons sont monophoniques.

6 Résultats

6.1 Premier ensemble de sons

Dans le tableau 1, sont présentés les résultats obtenus avec le système décrit dans [2] et [6], et ceux obtenus avec le système décrit dans cet article. La mise en place des méthodes à noyaux permet de réduire énormément le nombre de fausses alarmes : -79,3 %, ce au prix d'une dégradation très limitée du nombre de bonnes détections : -2,3 % seulement. La courbe COR donnée sur la figure 1 indique de plus que l'estimation automatique du seuil est correcte.

		référence [6]	
	nbre de ruptures	nbre de bd	nbre de fa
flûte	21	21	3
clarinette	30	29	2
violon	16	16	15
chant	21	21	9
total	88	87	29
		cet article	
	nbre de ruptures	nbre de bd	nbre de fa
flûte	21	20	1
clarinette	30	28	1
violon	16	16	3
chant	21	21	1
total	88	85	6

TAB. 1 – Les résultats obtenus dans [6] sont donnés dans le tableau du haut, et ceux obtenus avec le système décrit dans cet article sont donnés dans le tableau du bas ; « bd » indique « bonnes détections », et « fa » « fausses alarmes »

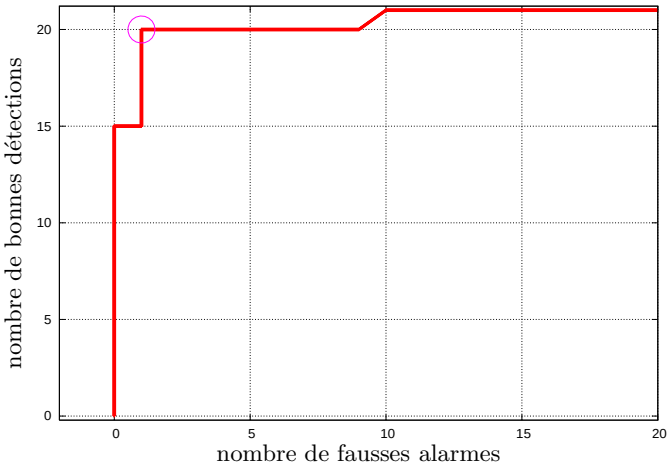

FIG. 1 – Courbe COR obtenue en utilisant le système décrit dans cet article pour l'extrait de flûte ; le cercle montre le résultat obtenu avec le seuil estimé automatiquement

En réglant les seuils pour les deux premiers sons afin que d'obtenir le même nombre de bonnes détections qu'avec le système de [6], nous obtenons en tout 26 fausses alarmes, soit une réduction de 10,3 % du nombre de fausses alarmes.

6.2 Second ensemble de sons

		référence [6]	
	nbre ruptures	nbre bd	nbre fa
gong 1	16	16	0
gong 2	16	16	0
violon	11	11	10
contrebasse	11	11	4
hautbois	8	8	9
clarinette	8	8	0
clarinette basse	8	8	5
contre-basson	8	8	4
saxophone 1	8	8	2
saxophone 2	8	8	4
grosse caisse	6	6	6
timbale	10	10	5
piano (Sqam)	11	11	20
accordéon	25	23	3
orgue	8	8	3
piano	2	2	3
total	164	162	78
		cet article	
	nbre ruptures	nbre bd	nbre fa
gong 1	16	16	1
gong 2	16	16	0
violon	11	11	0
contrebasse	11	10	0
hautbois	8	7	0
clarinette	8	8	0
clarinette basse	8	8	0
contre-basson	8	8	7
saxophone 1	8	7	0
saxophone 2	8	7	1
grosse caisse	6	6	3
timbale	10	10	1
piano (Sqam)	11	11	0
accordéon	25	24	0
orgue	8	7	0
piano	2	2	0
total	164	158	13

TAB. 2 – Les résultats obtenus dans [2] sont donnés dans le tableau du haut, et ceux obtenus avec le système décrit dans cet article sont donnés dans le tableau du bas ; « bd » indique « bonnes détections », et « fa » « fausses alarmes »

Dans le tableau 2, sont présentés les résultats obtenus avec le système décrit dans [2], et ceux obtenus avec le système décrit dans cet article, pour les sons du CD Sqam et l'extrait de piano supplémentaire. Pour ces sons aussi, la mise en place des méthodes à noyaux permet de réduire énormément le nombre de fausses alarmes : -83,3 %, ce au prix d'une dégradation très limitée du nombre de bonnes détections : -2,5 % seulement. En réglant les seuils pour obtenir le même nombre de bonnes détections qu'avec le système de [2], nous obtenons en tout 19 fausses alarmes, soit une réduction de 75,6 % du nombre de fausses alarmes.

7 Conclusion

Ces résultats de segmentation, obtenus sur un corpus assez vaste, montrent l'efficacité des méthodes à noyaux en ce qui concerne la détection de ruptures dans le cas de signaux sonores musicaux. Des résultats concernant l'optimisation des paramètres des diverses techniques utilisées, et des résultats concernant le choix optimal de l'ensemble des descripteurs audio à utiliser sont en cours d'obtention. Ce travail s'inscrit dans une recherche plus vaste, menée depuis plusieurs années par les auteurs. Cette recherche concerne la segmentation des sons en général.

Références

[1] Michèle Basseville et Igor V. Nikiforov, *Detection of abrupt changes*, PTR Prentice-Hall, 1993

[2] Stéphane Rossignol, *Segmentation et indexation des signaux sonores musicaux*, Université Pierre et Marie Curie, `http://stephanerossignol.ifrance.com`, Juillet 2000

[3] Frédéric Desobry, Manuel Davy et Christian Doncarli, *An Online Kernel Change Detection Algorithm*, IEEE Transactions on Signal Processing, Août 2005, pages 2961 – 2974. Volume 53, numéro 8.2

[4] Manuel Davy et S. J. Godsill, *Audio Information Retrieval : A Bibliographic Study*, Février 2002, CUED/F-INFENG/TR.429 ; Signal Processing Group, Cambridge University Engineering Department.

[5] Stéphane Rossignol, Peter Desain et Henkjan Honing, *State-of-the-art in fundamental frequency tracking : advantage of the availability of knowledge*, Proceedings of Workshop on Current Research Directions in Computer Music, Novembre 2001, pages 244 – 254

[6] Stéphane Rossignol, Xavier Rodet, Joël Soumagne, Jean-Luc Collette et Philippe Depalle, *Automatic characterisation of musical signals : feature extraction and temporal segmentation*, Journal of New Music Research (JNMR), Décembre 1999, volume 28, numéro 4, pages 281 – 295

[7] Mehmet Sezgin et Bülent Sankur, *Survey over image thresholding techniques and quantitative performance evaluation*, Journal of Electronic Imaging, Janvier 2004, volume 13(1), pages 146 – 165

[8] Philippe Depalle, Guillermo Garcia et Xavier Rodet, *Tracking of partials for additive sound synthesis using Hidden Markov Models*, IEEE Transactions on Acoustics, Speech, and Signal Processing (ICASSP93), 1993

[9] James A. Moorer, *On the segmentation and analysis of continuous musical sound*, PhD thesis, Stanford University, 1975