# Segmentation et indexation de documents sonores

S. Rossignol[1,2], J. Soumagne[2], X. Rodet[1], J.-L. Collette[2] and P. Depalle[1]

rossigno@ese-metz.fr,rod@ircam.fr,soumagne@ese-metz.fr

[1]IRCAM,1 place Igor-Stravinsky, 75004 PARIS, FRANCE

[2]Supélec 2, rue Édouard Belin, 57078 METZ, FRANCE

# 1 Introduction

La segmentation et l'indexation des signaux sonores musicaux est un domaine en pleine expansion, ceci du fait du développement de MPEG-7. Ce standard visera à définir un ensemble d'outils de description de contenus multimédias, pour faciliter la recherche et l'identification.

L'objectif que nous nous sommes fixé dans notre travail de thèse est de segmenter les sons musicaux en général en zones stables et d'étiqueter les segments obtenus. Pour un extrait de flûte, par exemple, il s'agit de segmenter en notes, puis de retrouver les notes jouées. Un son de flûte est considéré par nous comme étant un son « simple ». La segmentation et l'indexation de ces sons simples fait l'objet de la section 2 de l'exposé. Il apparaît des problèmes dès que le son n'est pas simple, c'est-à-dire dès qu'il n'est plus monophonique, ou harmonique, ou non-modulé. Des classes de sons sont définies. Il faut adapter le processus de segmentation à chacune d'elles. De plus, pour la voix chantée, nous pouvons vouloir segmenter en notes ou en phones. Ainsi, la notion de zone stable dépend du son, et aussi des désirs de l'utilisateur. Ces remarques nous ont amené à complexifier le processus de segmentation. Il se décompose en trois niveaux hiérarchiquement organisés ainsi : la segmentation en sources (section 4), la segmentation en caractéristiques (section 3) et la segmentation en zones stables (section 2). Les informations circulent du niveau le plus élevé au plus bas. Certains des liens entre les modules de segmentation sont détaillés dans la section 5. Nous concluons dans la section 6.

Ce travail de thèse est le résultat de la collaboration de trois centres de recherche : le CCETT, Supélec – Campus de Metz et l'IRCAM.

# 2 Segmentation d'un son monophonique, harmonique et non-modulé en zones stables (notes)

La procédure se décompose en quatre étapes. (1) Des fonctions d'observation sont définies qui réagissent au moment des transitions. (2) À partir de chaque fonction d'observation nous obtenons une fonction de décision. (3) Les fonctions de décision sont mises en concurrence (fusion de données hétérogènes). (4) L'indexation des segments finalement obtenus est effectuée.

## 2.1 Les fonctions d'observation

Il s'agit ici de mettre en évidence les variations brusques du signal (correspondant ici à des transitions entre notes). Trois types de variations brusques sont définis : les variations brusques de $f_0$, les variations brusques de l'énergie et les variations brusques du contenu spectral. Chaque fonction d'observation réagit plutôt à un seul type de variations, d'où déjà la nécessité d'utiliser plusieurs fonctions d'observation. En utilisant plusieurs fonctions d'observation, nous espérons de plus augmenter la robustesse du système : nous supposons que peu de fonctions d'observation ne réagissent pas au moment des transitions et que peu de fonctions d'observation réagissent à un instant donné au cours d'une zone stable. Nous attendons d'une fonction d'observation qu'elle présente un pic aussi fin et grand que possible au moment des transitions et un bruit de moyenne et de variance aussi faible que possible dans les zones stables.

### 2.1.1 Fonctions d'observation basées sur les variations de $f_0$

(1) Les dérivées numériques de $f_0$ sont utilisées. Les trajets de $f_0$ et des partiels sont obtenus grâce à un logiciel de l'IRCAM. (2) Ensuite, les coefficients de voisement sont définis. Nous corrélons des portions tronquées autour des $if_0$ et normalisées en énergie du spectre d'amplitude avec le spectre d'amplitude tronqué et normalisé de la fenêtre de pondération utilisée. La troncature est choisie telle que 90% de l'énergie du lobe de la fenêtre est conservée. Si une sinusoïde est présente en $if_0$, la corrélation est proche de 1. La moyenne des coefficients de voisement est proche de 1. Sinon, c'est-à-dire si $f_0$ n'a pas de sens, cette moyenne est petite. La fonction d'observation est l'inverse de la moyenne des dérivées des coefficients de voisement. Les spectres d'amplitude sont calculés sur des trames de quelques dizaines de ms. (3) Ensuite, nous avons considéré les indices d'inharmonicité. La position théorique du partiel $i$ est $if_0$ et sa position trouvée est $f_i$. L'indice d'inharmonicité pour le partiel $i$ est donc : $|if_0 - f_i|/(if_0)$. La moyenne des dérivées des indices d'inharmonicité est calculée. (4) Nous considérons ensuite l'écart statistique entre deux trames adjacentes. C'est la probabilité que l'écart entre 2 observations gaussiennes soit supérieur à un seuil S. S correspond ici à une donnée musicale : un écart de ton. (5) La dernière fonction d'observation est la rupture de modèle AR. Cette méthode est décrite dans (Basseville and Nikiforov).

### 2.1.2 Fonctions d'observation basées sur les variations de l'énergie.

Les dérivées, l'écart statistique et la rupture de modèle sont utilisées.

### 2.1.3 Fonctions d'observation basées sur les variations du contenu spectral.

(1) Les flux spectraux sont utilisés. Il s'agit de calculer : $\sum_{f=0}^{f=f_e/2} ||S_1(f)| - |S_2(f)||$, où $S_1$ et $S_2$ sont deux spectres d'amplitude ou deux enveloppes spectrales calculés pour 2 trames légèrement décalées. (2) Nous considérons ensuite un autre coefficient de voisement. Cette fois, nous ne retenons pas seulement les corrélations autour de $f_0$, mais les corrélations autour de tous les bins du spectre d'amplitude. Quand la corrélation dépasse un seuil s, une sinusoïde est détectée. Le rapport de la somme des énergies des sinusoïdes détectées et de l'énergie totale du signal nous indique si le signal est plutôt composé de sinusoïdes ou est plutôt du bruit.

## 2.2 Seuillages

Seules les fonctions d'observation basées sur la rupture de modèle AR nous donnent des fonctions de décision, c'est-à-dire des fonctions qui valent 0 dans les zones stables (bruit) et 1 lors des transitions (signal). Il faut seuiller les autres. Une vingtaine de méthodes de seuillage ont été testées. Celle qui est retenue pour le moment est la méthode des $3\sigma$. Les $N$ échantillons de la fonction d'observation sont rangés dans l'ordre croissant. Les $n$ plus petits sont conservés et nous permettent d'estimer $\sigma$, l'écart-type du bruit. Nous décidons ensuite que tous les pics supérieurs à $3\sigma$ correspondent à des pics à détecter. La méthode est robuste suivant $n$. Communément, $n$ est pris égal à 90%. Les seuils sont réactualisés toutes les 10 secondes. Il existe une justification théorique à ce seuil : voir le « seuil universel » de Donoho.

### 2.3 Prise de décision finale

Il faut ensuite fusionner les résultats obtenus avec toutes les fonctions d'observation. L'idée la plus simple est de sommer les fonctions de décision. Mais des problèmes apparaissent, du fait que les transitions ne sont pas instantanées, du fait que la fin d'une note se mélange souvent avec le début de la suivante (ceci à cause de la réverbération), etc. Ainsi, certaines fonctions d'observation réagissent plutôt au début de la transition, d'autres à la fin, et d'autres plusieurs fois au cours d'une transition. Quand nous sommons les fonctions de décision, nous obtenons pour chaque transition un nuage plus ou moins étalé de marques, et dans les cas extrêmes deux nuages de marques voisins se mélangent. Quelques techniques ont été testées qui nous permettent de concaténer les marques trop proches les unes des autres, de séparer les nuages de marques, etc.

### 2.4 Indexation

Il s'agit ici de retrouver les notes jouées. Pour cela, nous utilisons la moyenne et la médiane du trajet de $f_0$ entre deux marques de segmentation.

## 3 Segmentation en caractéristiques

Dès que le signal n'est plus monophonique ou harmonique ou non-modulé, la procédure décrite dans la section 2 ne suffit plus pour segmenter, notamment parce que certaines fonctions d'observation deviennent inutilisables. Par exemple, pour un signal de castagnettes, les variations de $f_0$ n'ont plus aucun sens, puisque le signal n'est pas voisé. Ce que nous voulons détecter ici, ce sont les transitoires d'énergie. Il faut avoir détecté que le signal n'est pas voisé avant d'essayer de segmenter en zones stables. C'est le but du niveau de segmentation en caractéristiques. Les caractéristiques considérées pour le moment sont : (1) présence ou absence de vibrato (modulation de fréquence) ; (2) présence ou absence de trémolo (modulation d'amplitude) ; (3) signal monophonique ou polyphonique ; (4) signal voisé ou non voisé (c'est-à-dire composé de sinusoïdes ou pas) ; (5) signal harmonique ou inharmonique (soit composé de sinusoïdes harmoniques ou pas) ; (6) signal ou silence.

Nous nous sommes particulièrement attaché au cours de cette thèse au problème du vibrato, c'est-à-dire à sa détection et à son extraction du trajet de $f_0$. Cette extraction est nécessaire, puisque l'excursion en fréquence d'un vibrato peut être supérieure au saut de fréquence entre deux notes : ainsi, les dérivées de $f_0$ ne peuvent qu'échouer à nous indiquer les transitions.

## 4 Segmentation en sources

Le but ici est de segmenter un son en parties voix parlée/parties musique instrumentale et/ou voix chantée. D'autres classes seront considérées par la suite.

### 4.1 Les fonctions d'observation

Elles sont calculées à partir des fonctions de base : (1) flux spectral sur deux trames légèrement décalées ; (2) centroïde (centre de gravité du spectre d'amplitude) sur une trame ; (3) taux de passage par 0 sur une trame dans le domaine temporel ; (4) flux entre le spectre d'amplitude et le cepstre liftré sur une trame. Et les fonctions d'observation sont la moyenne et le $\log_{10}$ de la variance de ces fonctions calculées sur un segment. Chaque trame fait 20 ms de long et chaque segment 1 seconde.

### 4.2 Les méthodes de classification

Les sons considérés sont deux sons de 20 mn chacun, formés de 10 mn de parole et de 10 mn de musique. Ces sons ont été enregistrés à la radio. 4 segments sur 5 servent à entraîner les classifieurs et le 5ème à les tester. Les 3 classifieurs utilisés sont : (1) La mixture de gaussiennes. Dans notre cas, nous avons modélisé chaque classe par une seule gaussienne multidimensionnelle. Il faudrait modéliser chaque classe par plusieurs gaussiennes, surtout la classe musique, qui est plus étendue, sa variance étant plus grande, que l'autre. (2) Les $k$ plus proches voisins. Pour chaque segment à classifier nous cherchons ses $k$ plus proches voisins (au sens de la distance euclidienne par exemple) dans les segments d'apprentissage. Le segment est rangé dans la classe la plus représentée dans ses $k$ voisins. (3) Les réseaux de neurones. Nous avons utilisé des réseaux de neurones à trois couches, aux unités complètement connectées. Nous avons considéré deux cas. Ou bien nous utilisons deux fonctions d'observation : alors la couche cachée comprend 6 unités. Ou bien nous utilisons les 6 premières fonctions d'observation : alors la couche cachée comprend 8 unités.

Les performances qui nous intéressent le plus sont les performances en validation croisée, c'est-à-dire les erreurs faites sur les données de test du son $i$ en utilisant les paramètres obtenus lors de l'apprentissage avec les données d'apprentissage du son $j$, avec $j \neq i$. Nous obtenons, avec le classifieur les 7 plus proches voisins : avec moy et var du flux spectral : 9,1% d'erreur ; avec moy et var du centroïde : 12,7% ; avec moy et var du taux de passage par 0 : 13,2% ; avec moy et var du flux spectre - cepstre liftré : 13,8%. Nous obtenons, avec les 6 premières fonctions d'observation, et les divers classifieurs : avec la mixture de gaussiennes : 9,3% ; avec les 7 plus proches voisins : 5,9% ; avec les réseaux de neurones : 4,7%.

## 5 Quelques liens entre les niveaux de segmentation

(1) Considérons le cas du vibrato. Quand de la parole est identifiée lors de la segmentation en sources, il n'est pas besoin d'essayer de détecter le vibrato. Quand les fonctions d'observation sont extraites pour le niveau de segmentation en zones stables, le vibrato n'est pas extrait du trajet de $f_0$. Si c'est de la musique qui est identifiée, nous essayons, lors de la segmentation en caractéristiques, de détecter le vibrato. S'il y en a un, il est extrait du trajet de $f_0$ et nous calculons les fonctions d'observation pour la segmentation en zones stables à partir du nouveau trajet de $f_0$.

(2) Le cas de la polyphonie. Quand la musique est identifiée lors de la segmentation en sources, le détecteur de polyphonie (nous décidons qu'un signal voisé mais inharmonique est polyphonique) est utilisé dans le niveau de segmentation en caractéristiques. Si la polyphonie est détectée, les fonctions d'observation basées sur trajet de $f_0$ ne sont pas utilisées pour la segmentation en zones stables.

## 6 Conclusion

Un système complet de segmentation a été défini et présenté. Il a été validé sur des signaux réels. Des programmes ont été développés en C.

Parmi les nombreuses perspectives possibles, nous pouvons retenir la séparation de sources, qui consiste à démixer des signaux mixés. La séparation de sources nécessite une segmentation performante.

## Références


[BN93] Michèle Basseville and Igor V. Nikiforov, *Detection of abrupt changes: Theory and application*, Prentice-Hall Inc., 1993.

[Don94] D. L. Donoho, *On minimum entropy segmentation*, Tech. Report 450, Stanford University, April 1994.

[RRS+00] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette, and P. Depalle, *Automatic characterisation of musical signals: feature extraction and temporal segmentation*, Journal of New Music Research, 2000.