Segmentation statistique de flux audio en temps-réel dans le cadre de la géométrie de l'information

Arnaud DESSEIN, Arshia CONT

IRCAM, CNRS UMR 9912
1 place Stravinsky, 75004 Paris, France
{dessein,cont}@ircam.fr

Résumé – Nous présentons un système temps-réel pour la segmentation statistique de flux audio dans le cadre de la géométrie de l'information. Le système repose sur le contrôle du rayon des boules de Bregman créées par les descripteurs observés et modélisés par une famille exponentielle donnée. Nous appliquons ce système à la segmentation spectrale de la musique polyphonique pour montrer la pertinence de l'approche proposée.

Abstract – We present a real-time system for statistical segmentation of audio streams in the framework of information geometry. The system relies on the control of the radius of the Bregman balls created by the observed descriptors that are modeled with a given exponential family. We apply this system to the spectral segmentation of polyphonic music to demonstrate the relevancy of the proposed approach.

1 Introduction

1.1 Motivations

Le problème de la segmentation audio a été largement étudié dans la littérature, notamment dans les domaines de la musique et de la parole [1, 2]. Ce problème consiste à partitionner un signal sonore en des régions temporelles continues homogènes, appelées segments, qui présentent des inhomogénéités avec les régions adjacentes. Cette définition met en relief les deux notions importantes de temporalité et d'homogénéité.

D'une part, la segmentation se doit de respecter la temporalité du signal sonore. De manière évidente, les segments formés doivent être continus le long de la dimension temporelle. De plus, l'information contenue dans un signal sonore est un flux qui se déroule de manière causale. Cette remarque prend tout son sens dans un contexte de segmentation en temps-réel, où l'on n'a pas accès au futur. La segmentation doit alors se faire de manière incrémentale, c'est-à-dire au fur et à mesure que le flux audio se déroule dans le temps, en comparant le présent au passé pour décider de la création d'un nouveau segment.

D'autre part, les segments doivent présenter une certaine homogénéité intrinsèque (une consistance de leur propre contenu informationnel) ainsi qu'une inhomogénéité avec les segments contigus (une différence de contenu informationnel avec les segments précédent et suivant). Cette notion pose donc les problèmes de définir un critère pour juger de l'homogénéité, et de quantifier l'information selon ce critère pour décider de la consistance ou non. Des critères très divers peuvent être utilisés suivant les types de signaux considérés. Il est par exemple possible de chercher à segmenter une conversation en termes de locuteurs présents, ou un programme radio en termes de passages d'émission, de musique et de publicité.

La plupart des travaux sur la segmentation audio repose sur des critères haut-niveau comme ceux mentionnés. En général, la segmentation repose sur une classification automatique qui engendre les segments en fonction des classes détectées (après un éventuel lissage temporel). En reprenant nos exemples, la segmentation d'une conversation dépendrait donc d'un système de reconnaissance du locuteur (et potentiellement d'un système de détection de silence). A l'identique, la segmentation d'un programme radio dépendrait d'un système de classification entre passages d'émission, de musique ou de publicité. Une telle approche a donc l'avantage de créer une segmentation haut-niveau, mais l'inconvénient de reposer sur un système de classification automatique qui n'est en général pas infaillible.

Nous aimerions pouvoir contourner ce problème, en segmentant un flux audio sans hypothèse sur l'existence de classes. Pour nous, la segmentation est donc en ça bien distincte de la classification ou du partitionnement de données, les classes étant remplacées par des unités consistantes d'un point de vue informationnel. Ces unités pouvant par la suite être considérées comme des symboles pour des traitements postérieurs, il serait également appréciable que l'on puisse caractériser chaque unité par un élément représentatif. Il s'agit donc d'une sorte de quantification du flux audio au cours de son évolution temporelle, mais sans chercher a priori à relier les unités quantifiées entre elles ou à leur attribuer une quelconque classe.

Des approches n'utilisant pas de classification automatique ont déjà été abordées, en particulier pour le problème spécifique de segmentation de la parole en locuteurs [3, 4]. Les méthodes proposées consistent en général à calculer une distance entre des trames successives, ou une statistique sur l'hypothèse d'un changement aux différentes trames, afin de décider de la création d'un nouveau segment.

1.2 Contributions

Nous proposons de formuler le problème de la segmentation audio d'un point de vue statistique dans le cadre de la géométrie de l'information. Dans ce cadre théorique, nous développons un système générique pouvant fonctionner pour différents types de signaux et de critères d'homogénéité que l'utilisateur définit selon l'application souhaitée.

De manière générale, le système segmente un flux audio en contrôlant la variation d'information contenue dans le flux. En d'autres termes, plutôt que d'analyser le signal bas-niveau pour prédire une classe haut-niveau qui le caractérise, nous considérons directement des changements de l'information contenue dans le signal. Les signaux et critères d'homogénéité pouvant prendre des formes très différentes, nous nous tournons vers des mesures probabilistes sur des descriptions statistiques du signal. Les unités quantifiées sont ensuite représentées par des modèles probabilistes susceptibles de servir d'entités abstraites symboliques pour des traitements postérieurs.

Plus en détail, à partir de la forme d'onde du flux audio arrivant, l'idée est de calculer des descripteurs à court-terme puis de modéliser ces descripteurs par des densités de probabilité. Les descripteurs audio ainsi que leurs modèles probabilistes associés représentent les critères d'homogénéité pour la segmentation et sont donc laissés au choix de l'utilisateur dans une certaine mesure. En particulier, nous considérons des densités issues de familles exponentielles. La géométrie de l'information fournit un cadre théorique riche pour étudier la structure géométrique intrinsèque de ces familles et développer des méthodes computationnelles génériques afin de les manipuler d'un point de vue informationnel.

2 Système proposé

2.1 Cadre de la géométrie de l'information

La géométrie de l'information est un champ des mathématiques qui étudie les notions de probabilité et d'information par le biais de la géométrie différentielle. L'idée principale est d'analyser la structure géométrique de variété différentielle que possèdent certaines familles paramétriques de densités de probabilité $\mathcal{S} = \{p_\xi : \xi \in \Xi\}$. Les travaux fondateurs dans ce domaine sont attribués à Rao qui a introduit, sous certaines hypothèses, la matrice d'information de Fisher comme métrique riemannienne sur une variété statistique, lui conférant ainsi une structure de variété riemannienne [5]. Par la suite, Chentsov a formalisé ce cadre et montré que, sous certaines conditions, la métrique d'information de Fisher est l'unique métrique riemannienne sur une variété statistique [6]. Amari et Nagaoka ont ensuite introduit une notion de dualité entre les connexions affines alpha $\nabla^{(\alpha)}$ et $\nabla^{(-\alpha)}$ sur une variété statistique, ce qui a mené à la formulation moderne du cadre de la géométrie de l'information [7].

Nous nous concentrons sur les variétés statistiques liées aux familles exponentielles. Les familles exponentielles considérées ici sont des familles paramétriques de densités de probabilité qui s'écrivent sous la forme canonique suivante :

$$p_\theta(x) = \exp\left(\theta^\top T(x) - F(\theta) + C(x)\right) \quad \forall\, x \in \mathcal{X}\,, \qquad (1)$$

où le vecteur θ représente les paramètres naturels et appartient à un ouvert convexe $\Theta \subset \mathbb{R}^n$, la fonction $F\colon \Theta \to \mathbb{R}$ est strictement convexe et infiniment différentiable, les fonctions $C\colon \mathcal{X} \to \mathbb{R}$ et $T\colon \mathcal{X} \to \mathbb{R}^n$ sont mesurables. Les familles exponentielles regroupent la grande majorité des distributions discrètes et continues employées en apprentissage statistique (normale, exponentielle, Poisson, Bernoulli, binomiale, catégorique, multinomiale, etc.).

Il est connu qu'une famille exponentielle $\mathcal{S} = \{p_\theta : \theta \in \Theta\}$ est une variété statistique sur laquelle la métrique d'information de Fisher g est l'unique métrique riemannienne vérifiant certaines propriétés d'invariance. De plus, la variété statistique $(\mathcal{S}, g, \nabla^{(1)}, \nabla^{(-1)})$ possède deux systèmes de coordonnées affines duaux, respectivement les paramètres naturels $\theta \in \Theta$ et les paramètres d'espérance $\eta \in \mathrm{H}$, avec lesquels il est pratique de travailler d'un point de vue computationnel, et qui sont reliés par la relation $\eta = \nabla F(\theta)$. La géométrie dualement plate sous-jacente exhibe une structure duale hessienne, engendrée par le potentiel F et son potentiel conjugué $F^\star$. Ce potentiel conjugué est défini par la transformation de Legendre-Fenchel :

$$F^\star(\eta) = \sup_{\theta \in \Theta} \theta^\top \eta - F(\theta) \quad \forall\, \eta \in \mathrm{H}\,, \qquad (2)$$

et vérifie $\nabla F^\star = (\nabla F)^{-1}$ de sorte que $\theta = \nabla F^\star(\eta)$. Une telle géométrie généralise la géométrie euclidienne auto-duale, avec notamment deux divergences de Bregman $\mathcal{B}_F$ et $\mathcal{B}_{F^\star}$ en lieu et place de la distance euclidienne auto-duale, mais aussi des géodésiques duales, un théorème de Pythagore généralisé et des projections duales.

Sur les variétés statistiques considérées, la notion de distance est donc définie par des divergences de Bregman. La divergence de Bregman $\mathcal{B}_G$, générée par une fonction $G\colon \Xi \to \mathbb{R}$ convexe et différentiable, entre deux points $\xi, \xi' \in \Xi$ est définie de la manière suivante :

$$\mathcal{B}_G(\xi \parallel \xi') = G(\xi) - G(\xi') - (\xi - \xi')^\top \nabla G(\xi')\,. \qquad (3)$$

Les divergences de Bregman sont des distances généralisées : elles sont positives et s'annulent ssi $\xi = \xi'$, mais elles ne sont pas symétriques et ne vérifient pas l'inégalité triangulaire en général. En revanche, les divergences de Bregman duales $\mathcal{B}_F$ et $\mathcal{B}_{F^\star}$ associées à une famille exponentielle sont pertinentes d'un point de vue statistique et informationnel car elles sont liées à la divergence de Kullback-Leibler (elle-même liée localement à la métrique d'information de Fisher) par la relation suivante :

$$\mathcal{D}_{\mathrm{KL}}(p_\xi \parallel p_{\xi'}) = \mathcal{B}_F(\theta' \parallel \theta) = \mathcal{B}_{F^\star}(\eta \parallel \eta')\,, \qquad (4)$$

où θ, θ' et η, η' sont respectivement les paramètres naturels et d'espérance des densités $p_\xi, p_{\xi'}$. C'est donc à l'aide de ces divergences informationnelles que nous formulons le problème de segmentation.

2.2 Formulation du problème de segmentation

Nous considérons un flux audio qui arrive de manière incrémentale au système sous forme de trames successives x_j. Chaque trame x_j est d'abord représentée par un descripteur à court-terme d_j (spectre d'amplitude, etc.). Les descripteurs sont ensuite modélisés par des densités de probabilité p_{θ_j} issues d'une famille exponentielle $\mathcal{S} = \{p_\theta : \theta \in \Theta\}$ donnée. Le choix des descripteurs et de la famille est laissé à l'utilisateur en fonction des signaux et critères d'homogénéité considérés. Le cadre de la géométrie de l'information permet alors de proposer un paradigme de segmentation indépendant de ce choix.

Au fur et à mesure que les points θ_j arrivent, nous les agrégeons dans une boule de Bregman. Comme les divergences de Bregman sont asymétriques en général, une boule de Bregman peut être considérée soit à gauche soit à droite. Par dualité, il est évident de remarquer qu'une boule de Bregman à droite sur les paramètres naturels correspond à une boule de Bregman à gauche sur les paramètres d'espérance et réciproquement. Nous nous restreignons donc sans perte de généralité au cas d'une boule de Bregman à droite sur les paramètres naturels, définie par un centre $\hat{\theta}$ et un rayon r comme suit :

$$B_F(\hat{\theta}, r) = \{\theta \in \Theta : \mathcal{B}_F(\theta \parallel \hat{\theta}) \leq r\}\,. \tag{5}$$

Durant l'agrégation des points, le centre $\hat{\theta}$ de la boule est mis à jour de manière incrémentale à l'arrivée de chaque nouveau point. Ce centre est calculé comme le centre de gravité à droite des points $\theta_1, \ldots, \theta_n \in \Theta$ déjà arrivés :

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{j=1}^{n} \mathcal{B}_F(\theta_j \parallel \theta)\,. \tag{6}$$

Ce problème d'optimisation n'est en général pas convexe, mais admet néanmoins l'unique solution globale suivante :

$$\hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} \theta_j\,. \tag{7}$$

En ce qui concerne le rayon, celui-ci est également mis à jour après le calcul du nouveau centre comme maximum des divergences entre les points de la boule et le centre :

$$r = \max_{j \in \{1,\ldots,n\}} \mathcal{B}_F(\theta_j \parallel \hat{\theta})\,. \tag{8}$$

Lorsque le rayon r de la boule devient supérieur à un certain seuil γ, nous segmentons le flux audio. Plus précisément, nous n'ajoutons pas le point en cours à la boule mais nous créons une nouvelle boule avec ce point comme premier point. L'ancienne boule n'est plus modifiée par la suite ; les points qui la composent déterminent un segment. Le schéma de segmentation mentionné est alors répété du début à partir du point en cours jusqu'à la création d'une nouvelle boule et ainsi de suite.

L'interprétation statistique et informationnelle de ce schéma géométrique de segmentation est la suivante. Les boules de Bregman déterminent des unités dans lesquelles les points ont un contenu informationnel consistant par rapport à un seuil d'information donné. Les trames correspondantes du flux audio forment un segment, et ont une certaine homogénéité intrinsèque d'un point de vue statistique pour le critère désiré. Les points respectifs de deux boules successives contiennent en revanche une information différente, ce qui implique une certaine inhomogénéité entre les trames respectives des segments correspondants. Pour finir, les centres des boules de Bregman sont des candidats idéaux pour représenter l'information contenue dans chaque segment, et servir d'entités abstraites symboliques pour des traitements postérieurs.

3 Résultats expérimentaux

3.1 Protocole

Nous illustrons ici le système de segmentation temps-réel proposé pour le cas de la musique polyphonique. L'exemple considéré est un passage du 1er mouvement *Pavane de la Belle au bois dormant* extrait de *Ma mère l'Oye, Cinq pièces enfantines pour piano à quatre mains* (1908-1910) par Maurice Ravel (1875-1937). Cet extrait a été synthétisé à partir d'une partition MIDI en utilisant des échantillons de piano réels.

Nous avons choisi pour cet exemple un critère d'homogénéité spectral, c'est-à-dire que nous avons voulu segmenter le passage musical au niveau des variations du contenu fréquentiel. Les trames entrant au système ont été représentées par leur spectre d'amplitude normalisé calculé par une simple transformée de Fourier à court-terme avec des trames de 512 échantillons et un pas d'avancement de 128 échantillons à une fréquence d'échantillonnage de $11025\,\text{Hz}$. Le modèle statistique choisi est celui des distributions catégoriques, déjà employées en traitement de l'image [8] et de l'audio [9]. Le schéma de segmentation a été considéré à droite sur les paramètres d'espérance et le seuil de segmentation a été fixé à $\gamma = 0.8$. La segmentation a été calculée sous MATLAB avec un ordinateur portable $2,40\,\text{GHz}$ / $4,00\,\text{Go}$ RAM, et a été environ 10 fois plus rapide que le temps-réel.

3.2 Discussion

Le résultat de la segmentation est représenté Figure 1 par des traits verticaux sur la forme d'onde et sur la partition MIDI de l'extrait musical. De manière générale, le système proposé a été capable de segmenter le flux audio en tranches polyphoniques stationnaires : les segments sont compris entre des débuts et/ou fins de notes. Nous insistons sur le fait que le système n'a pourtant aucune connaissance sur ce qu'est une note de musique, et que la segmentation se fonde uniquement sur la variation d'information, en l'occurrence spectrale, entre les trames successives du flux audio. De plus, en utilisant un spectre normalisé, nous n'avons pas pris en compte de notion d'énergie sonore mais seulement l'amplitude relative des différentes fréquences, ce qui constitue une alternative intéressante aux systèmes de détection de début de notes qui reposent pour la plupart sur des critères d'énergie.

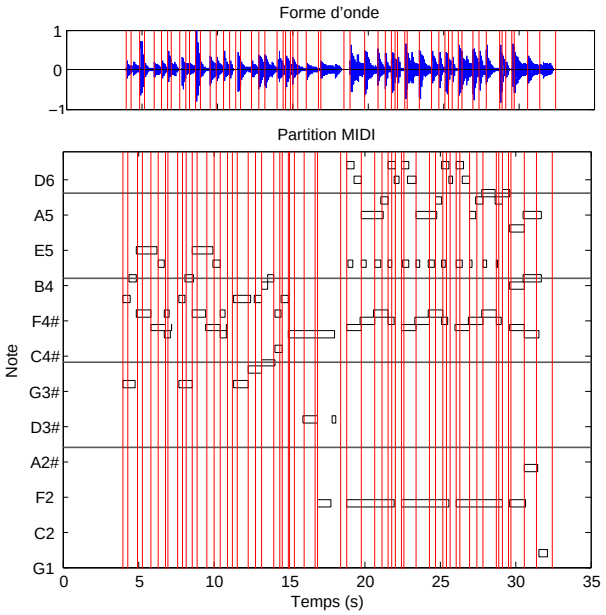

FIGURE 1 – Segmentation d'un extrait musical polyphonique.

Nous notons toutefois que certaines notes sont également segmentées après l'attaque, lorsque celle-ci possède un contenu fréquentiel trop différent du reste de la note dû à la présence non négligeable de transitoires. Remarquons aussi que certains débuts et fins de notes ne sont pas segmentés, notamment quand des notes jouées simultanément sont en rapport harmonique fort et ont donc de nombreuses fréquences communes.

En prenant en compte ces remarques, le système pourrait être amélioré en fonction de l'application souhaitée. Pour réaliser un détecteur de début de notes, on pourrait par exemple ajouter de l'information sur l'énergie. Pour employer la segmentation comme un module de prétraitement dans un système de transcription automatique, on pourrait forcer une sur-segmentation en diminuant le seuil, afin d'avoir uniquement des tranches polyphoniques stationnaires, et estimer les hauteurs présentes dans ces tranches en utilisant leurs centres gravités respectifs.

4 Conclusion

Nous avons présenté un système temps-réel pour la segmentation statistique de flux audio dans le cadre de la géométrie de l'information. Nous avons montré la pertinence de ce système pour la segmentation spectrale de la musique polyphonique.

Dans notre approche, nous avons directement modélisé les descripteurs utilisés par des densités issues d'une famille exponentielle. Nous voulons étendre notre cadre de segmentation au cas où les descripteurs sont des observations indépendantes distribuées selon des densités d'une famille exponentielle. Dans ce cas, il est possible de considérer notre schéma de segmentation sur les statistiques exhaustives des observations. Nous pensons cependant qu'une telle approche ne serait pas suffisamment robuste, et qu'il faudrait alors plutôt utiliser des tests séquentiels sur l'hypothèse d'un changement aux différentes trames [10].

Enfin, nous souhaitons évaluer le système proposé de manière exhaustive dans des applications telles que la détection de débuts de notes et la segmentation de la parole en locuteurs.

Remerciements

Ce travail a été en partie financé par une allocation doctorale de l'UPMC (EDITE) et une bourse du JST-CNRS ICT (Improving the VR Experience).

Références

[1] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1423–1426, Istanbul, Turquie, Juin 2000.

[2] H. Sundaram and S.-F. Chang. Audio scene segmentation using multiple features, models and time scales. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 2441–2444, Istanbul, Turquie, Juin 2000.

[3] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 97–99, Chantilly, VA, USA, Février 1997.

[4] A. Tritschler and R. A. Gopinath. Improved speaker segmentation and segments clustering using the Bayesian information criterion. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, pages 679–682, Budapest, Hongrie, Septembre 1999.

[5] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37 :81–91, 1945.

[6] N. N. Chentsov. *Statistical decision rules and optimal inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, 1982.

[7] S.-i. Amari and H. Nagaoka. *Methods of information geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[8] F. Nielsen and R. Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6) :2882–2904, Juin 2009.

[9] A. Cont, S. Dubnov, and G. Assayag. On the information geometry of audio streams with applications to similarity computing. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4) :837–846, Mai 2011.

[10] M. Basseville and I. V. Nikiforov. *Detection of abrupt changes: Theory and application*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.