

Suivi d'évènements sonores multiples dans les documents audiovisuels

M. Betser¹

G. Gravier¹

¹ IRISA (INRIA & CNRS) / METISS
Campus Universitaire de Beaulieu, Rennes

{mbetser, ggravier}@irisa.fr

Résumé

Détecter et suivre des classes de sons dans un document sonore est une étape importante pour la structuration du contenu. Dans le cas de scènes sonores complexes, comme des bandes sons télévisuelles, la détection d'évènements audio simultanés est un problème encore mal résolu. Dans cet article, une approche en deux étapes est proposée pour détecter ces évènements superposés. La première consiste en une segmentation aveugle, suivie d'une étape de détection sur chaque segment. Afin de mieux évaluer la qualité du système, de nouvelles mesures de performances sont introduites, plus appropriée cette tâche. Une approche Viterbi équivalente est également développée à titre de comparaison.

Mots clefs

Suivi, Évènements sonores, MAP, Viterbi, Tennis.

1 Introduction

Il existe deux grands types d'approches pour la détection et le suivi d'évènements dans un document audio. La première consiste à le segmenter en plages acoustiquement homogènes, avec une éventuelle phase de regroupement, avant une étape de détection. Cette dernière va étiqueter les différents segments, ou groupes de segments, avec les évènements acoustiques correspondant ou bien va y chercher un évènement particulier [1]. Dans cette approche en deux étapes, la segmentation repose généralement sur un critère d'information, et la classification utilise des modèles statistiques dont les paramètres sont estimés sur des corpus d'entraînement.

L'autre type d'approche réalise une segmentation et une classification conjointe à partir de modèles statistiques. Elle est beaucoup utilisée dans le cadre du suivi de locuteur par modèles de mélange de Gaussiennes (MMG) [2]. Comme nous l'avons souligné dans un précédent article [3], la plupart des systèmes supposent que les évènements sonores ne peuvent avoir lieu simultanément, alors que la plupart des documents, particulièrement les documents sportifs, ne peuvent être représentés comme une succession d'évènements isolés. Afin de pouvoir les décrire complètement, il serait donc souhaitable de disposer de méthodes génériques pour détecter des évènements audio

simultanés. Quelques méthodes ont été proposées dans [3] pour réaliser une telle détection.

Cet article s'inscrit dans le cheminement de notre précédent travail et s'intéresse plus particulièrement à l'approche en deux étapes. L'étape de détection proposée sépare le problème global en un certain nombre de problèmes classe vs non-classe indépendants au sein d'une approche de type maximum a posteriori (MAP). De nouvelles mesures de performances adaptées à la tâche de détection d'évènements simultanés sont proposées. Les expériences sont réalisées sur un corpus de vidéo de tennis. Nous étudions plus particulièrement l'influence de l'étape de segmentation et d'approximations sur les probabilités a priori des évènements. Enfin, nous tentons d'appliquer les principes de cette détection au problème de segmentation et de classification conjointe, en utilisant des modèles de Markov cachés (MMC).

2 Une approche de type MAP pour le suivi d'évènements sonores

Le but de ce travail¹ est la détection, dans un document audio, d'évènements sonores, définis arbitrairement, pouvant apparaître simultanément. Nous nous intéressons plus particulièrement à une détection en deux étapes, avec un découpage a priori en segments homogènes (homogène signifie ici que les classes sonores présentes le sont sur tout le segment), une éventuelle étape de regroupement des segments similaires non-adjacents, et une classification réalisée à l'aide de MMGs.

Dans un travail antérieur, plusieurs méthodes ont été proposées, la meilleure étant une approche Viterbi, où les modèles d'évènements superposés étaient calculés par concaténation de modèles d'évènement simple (méthode de concaténation). Une première détection en deux étapes fut également testée, mais sans apporter d'amélioration. L'étape de segmentation était basée sur un critère d'information Bayésien (CIB), sans regroupement, et la détection sur un test statistique à deux hypothèses pour chaque classe sonore, celles-ci pouvant être *présentes* ou *non présentes*. Les modèles de présence de classe étaient estimés sur des

¹Ce travail a été réalisé dans le cadre du projet Domus Videum mis en place par le RNRT

segments ou seul la classe considérée était présente, et le modèle de non présence sur tous les segments qui ne contenaient pas cet évènement.

Nous décrivons ici les trois étapes, de segmentation, de regroupement et de détection, de notre approche.

Segmentation et regroupement

Comme dans notre précédent système en deux étapes, la segmentation utilise un CIB. Les résultats présentés sont issus d'un algorithme en trois passes: d'abord une détection grossière des changements avec une fenêtre grandissante; ensuite un affinement des frontières; et enfin validation des ruptures. Un algorithme en deux étapes utilisant une courbe indicatrice de variation de CIB a aussi été testé, mais les résultats étant très similaires, nous avons choisis de ne pas les présenter.

L'algorithme de regroupement utilisé est de type hiérarchique ([4]). Chaque groupe est initialisé avec un segment. A chaque étape, une matrice de distance entre groupes est calculée en utilisant la mesure de similarité de Kullback-Liebler (KL). Deux segments ne pouvant être regroupés seulement en cas d'augmentation d'un CIB global, la deuxième étape consiste à trouver les deux groupes de segments les plus proches (au sens KL) vérifiant ce critère, et de les regrouper.

Les calculs de similarités KL et du CIB se font, comme dans l'étape de segmentation, en modélisant les groupes, ou les segments, par de simples Gaussiennes.

Classification

Une façon d'aborder la détection d'évènement est de considérer chaque évènement comme une source, qui émet ou non. Le signal peut alors être représentée comme un mélange de ces sources, et chaque segment par un état $X = \{X_i, i = 1..d\}$, où X_i est égal à un si l'évènement i est présent ou à zéro dans le cas contraire, d étant le nombre d'évènements considérés.

Le critère MAP est utilisé pour déterminer l'état le plus probable \hat{x} parmi tous les états possibles $x = \{x_i = (0, 1), i = 1..d\}$,

$$\hat{x} = \arg \max_x P(X = x|y) \quad (1)$$

ou y représente les données observées d'un segment. Le problème est d'évaluer le vraisemblance de l'observation $P(y|X)$ pour chaque état possible ainsi que les probabilités à priori correspondante $P(X)$. Si plus de deux évènements sont présents, il devient impossible d'estimer tous ces modèles par manque de données d'entraînement. Nous proposons donc de supposer l'indépendance statistique des classes vis à vis des données. La vraisemblance d'une observation y pour un état X peut être calculée de la façon suivante:

$$\begin{aligned} P(X = x|y) &= \prod_i P(X_i = x_i|y) \\ &\sim \prod_i P(y|X_i = x_i) \cdot P(X_i = x_i) \end{aligned} \quad (2)$$

Dans ce cas de figure, seuls deux modèles $P(y|X_i = 1)$ et $P(y|X_i = 0)$ sont nécessaires. Remarquons que dans ce formalisme, $P(y|X_i = 1)$ représente toutes les données où la classe i est présente. Elle est différente de $P(y|X_i = 1, X_{k \neq i} = 0)$ qui représente les données où la classe est présente seule, fait vérifié expérimentalement. Le critère MAP peut être ré-écrit ainsi:

$$\hat{x} = \arg \max_x \prod_{i, x_i=1} \frac{P(X_i = 1)}{P(X_i = 0)} \cdot \frac{P(y|X_i = 1)}{P(y|X_i = 0)} \quad (3)$$

Il est donc équivalent, de prendre des décisions indépendamment pour chaque évènement en comparant le rapport de vraisemblance $l_i = P(y|X_i = 1)/P(y|X_i = 0)$ au seuil $\beta_i = P(X_i = 0)/P(X_i = 1)$. Il apparaît clairement que si $l_i < \beta_i$, l'état \hat{x}_i doit être égal à zéro pour maximiser (3).

Dans un précédent article, un seuil unique a été utilisé pour toutes les classes, ce qui correspond au cas où toutes les sources ont la même probabilité à priori. Comme nous le verrons dans la section 4.1, plusieurs approximations des probabilités à priori ont été essayées.

3 Mesure de performance

Ce travail se place du point de vue de la détection d'évènements, où les fausses alarmes et les faux rejets ont un coût égal.

Soit $T_i(1, 1)$ et $T_i(0, 0)$ les durées totales correctement détectées pour les évènements présence de i , et absence de i . De la même manière, $T_i(1, 0)$ et $T_i(0, 1)$ sont respectivement les temps totaux de fausse alarme et de détection manquée. Enfin, $T_i(1)$ et $T_i(0)$ sont les durées totales dans la référence respectivement de la présence de i et de son absence. En utilisant les notations précédentes, les mesures de performances utilisées le cadre de la détection peuvent être étendues pour l'étude d'évènements superposés. Nous définissons les trois mesures suivantes:

$$\begin{aligned} \% \text{corr} &= \frac{\sum_i T_i(1, 1) + T_i(0, 0)}{d.T} \\ \% \text{FA} &= \frac{\sum_i T_i(0, 1)}{\sum_i T_i(1)} \\ \% \text{FR} &= \frac{\sum_i T_i(1, 0)}{\sum_i T_i(0)} \end{aligned}$$

$\% \text{corr}$ est le taux de classification moyen parmi tous les évènements considérés. On peut facilement vérifier que:

$$\% \text{corr} + \frac{\sum_i T_i(1)}{d.T} \cdot \% \text{FA} + \frac{\sum_i T_i(0)}{d.T} \cdot \% \text{FR} = 1$$

Comme notre but est de détecter des évènements superposés, nous avons également besoin d'un indicateur pour mesurer la qualité de la détection des évènements multiples. Le taux de reconnaissance des segments à évènements multiples, $\% \text{mcorr}$, est défini comme le rapport entre la durée des segments à évènements multiples correctement reconnus et la durée totale des segments à évènements multiples dans la segmentation de référence.

	<i>prior1</i>	<i>prior2</i>	<i>prior3</i>	<i>Ref.</i>	<i>Davis</i>	<i>Vit. 2</i>	<i>Vit. 1</i>
% corr	89	89	88	93	92	89	88
% FA	22	16	26	13	13	25	20
% FR	7	9	6	5	7	6	9
% mcorr	54	34	54	65	63	54	32

Tab. 1 – Les 5 premières colonnes sont les résultats des tests d’hypothèse binaires: les trois types d’à priori testés dans la section 2, une expérience réalisée sur la segmentation de référence, et un test réalisé sur la fin du match de la coupe Davis uniquement. ‘Vit. 2’ donne les résultats pour la méthode des MMCs à 2 états de la section 5, et ‘Vit. 1’ rappelle ceux du Viterbi avec concaténation des modèles.

4 Expériences

Les expériences sont réalisées sur trois vidéo de tennis. Le corpus d’entraînement contient un match complet (Roland Garros) et deux premiers sets d’un match de la coupe Davis. Le dernier set de ce match, plus trois autres tirés d’un tournoi de Bercy composent le corpus de test. La tâche considérée est la structuration en quatre classes sonores, *parole, musique, applaudissements, tennis*, des bandes son. Cette dernière classe correspond aux bruits des phases de jeu comme les cris des joueurs, les rebonds de balles, etc... On estime un MMG à 64 composantes et à matrices de covariance diagonales par modèle. Le signal est représenté par des coefficients cepstraux avec les dérivées du premier et du second ordre. Les résultats sont résumés dans le tableau 1.

4.1 Approximation des probabilités à priori

Dans cette partie, nous comparons plusieurs approximation des probabilités à priori, $P(X_i = 0)$ et $P(X_i = 1)$. Elles peuvent être estimées sur le corpus d’entraînement (*prior1*). En faisant l’hypothèse que les sources ont les même probabilités à priori, un seuil optimal global peut également être déterminé sur le corpus d’entraînement (*prior2*). Lorsque la décision est prise indépendamment pour pour chaque évènement, la pénalisation des états où beaucoup d’évènements sont présents est moins marquée, alors que dans le cas du tennis, par exemple, il y aura rarement plus de deux classes présentes en même temps. Nous avons donc essayé un à priori donnant une probabilité égale aux états avec zéro, une ou deux classes présentes et une probabilité de zéro aux autres états (*prior3*).

Les meilleurs résultats sont obtenus pour le premier type d’à priori, les autres dégradant légèrement les performances. Néanmoins, lorsque les probabilités à priori exactes ne sont pas disponibles, des approximations peuvent être utilisées sans diminution dramatique des performances. Notons également que ces trois à priori donnent des types d’erreurs différents. Suivant le type d’erreur à minimiser, un à priori différent peut être envisagé.

Par rapport à notre système de Viterbi de référence (Vit. 1), l’approche proposée conduit à de meilleurs résultats, en particulier pour la détection des évènements superposés. Si la structuration est réalisée avec les même conditions acoustiques que pour le corpus d’entraînement, on constate

une bonne amélioration des performances (Davis), ce qui suggère qu’une adaptation des modèles sur les données de test peut se révéler judicieuse.

4.2 Influence de la segmentation

Le paramètre théorique (λ) du CIB est 1, mais comme le montre la figure 1, il induit des segments de taille importante, plus grands qu’ils ne devraient être (la longueur moyenne des segments pour la référence est de 4,99s). Lorsque l’on diminue le seuil, on détecte plus de segments. La segmentation donnant les meilleurs résultats, pour un paramètre CIB de 0.6, a approximativement le même nombre de segments que la référence. Cette valeur de λ a été retenue pour toutes les autres expériences. Cependant, on constate toujours une nette baisse des performances par rapport à une segmentation parfaite (manuelle). Cela signifie, que certains changements ont été mal placés, et par conséquent que certains évènements que nous désirons détecter ont été ignorés. Ces évènements manquants, ne sont toujours pas détectés pour de faibles valeurs du paramètre CIB.

Le même problème se pose avec l’algorithme de regroupement. On ne constate pas d’amélioration après cette étape: d’un côté, l’estimation des classes présentes est meilleure globalement pour les regroupements que pour les segments non regroupés, mais d’un autre côté, certains segments avec des évènements sonores superposés vont être placés dans de mauvais regroupements, en particulier lorsqu’il y a un évènement fortement dominant (typiquement lorsqu’il y a des commentaires sur du tennis).

Le modèle mono-gaussien utilisé pour comparer les segments est une approximation trop grossière dans le cas d’évènements sonores superposés. Les taux de classification obtenus pour une segmentation parfaite (Ref.) suggèrent que des modèles de classes multi-gaussiens bien estimés permettent de mieux distinguer des évènements superposés. Il est donc raisonnable de penser que leur intégration dans le processus de segmentation devrait améliorer le système.

5 MMCs à deux états

Une façon d’introduire les modèles de classe dans la phase de segmentation consiste à utiliser un algorithme Viterbi.

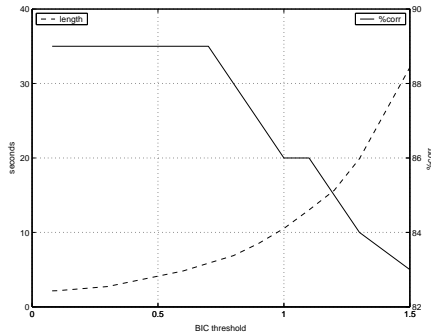


Fig. 1 – La première courbe (en trait plein) est le %corr et la seconde courbe (en pointillé) la longueur moyenne des segments, en fonction du paramètre λ du CIB.

Dans cet algorithme, les phases de segmentation et de classification sont mélangées. Afin de pouvoir garder la représentation classe vs non-classe qui est l'idée central de cet article, nous proposons une approche Viterbi utilisant deux modèles d'état pour chaque classe.

Si l'on considère un MMC avec tous les états globaux possibles (cf. figure 2(a)), l'algorithme de Viterbi assure que le chemin trouvé sera optimal au sens du maximum de vraisemblance. Si nous faisons les mêmes hypothèses d'indépendance que dans la deuxième section, les vraisemblances d'un état global peuvent s'exprimer comme le produit de vraisemblances de classe et de non-classe (2). A partir de là, on peut montrer facilement que l'utilisation de ce MMC est équivalente à celle de d MMC à deux états, un pour chaque classe (cf. figure 2(b)).

En résumé, l'utilisation d'un algorithme Viterbi pour d MMC à deux états classe et non-classe, et en concaténant les segmentations résultantes, conduit à une segmentation globalement optimale sous les deux hypothèses d'indépendances de la deuxième section.

Expériences

La modélisation classe vs non-classe utilisée dans les MMC à deux états est la même que celle utilisée dans les tests d'hypothèse. Les probabilités de transition sont estimées sur le corpus d'entraînement.

Les résultats des MMC à deux états (Vit. 2) sont équivalents à ceux obtenus avec les tests d'hypothèses. Cette modélisation est donc une alternative intéressante à la segmentation aveugle par CIB. Les résultats ne sont pas meilleurs car certaines hypothèses structurelle du Viterbi altèrent aussi la segmentation. En effet la classification de chaque trame ne dépend que de la précédente (hypothèse Markovienne d'ordre 1), ce qui représente très imparfaitement la notion de segment. Pour éviter de trop fréquents changements d'états dans les MMCs, nous avons utilisé une technique classique consistant à pénaliser les transitions entre deux états différents. Dans ce cas, il y a des imprécisions sur les transitions des segments et les petits segments seront absorbés par les grands segments voisins.

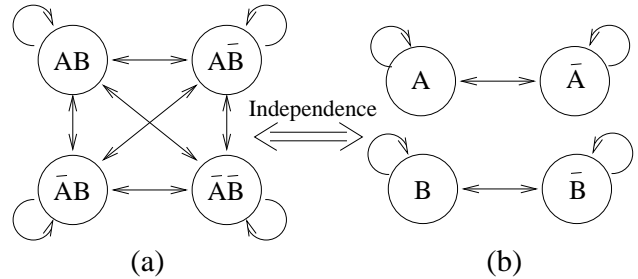


Fig. 2 – Pour deux classes sonores indépendantes A et B , l'équivalence entre un MMC avec tous les états globaux possibles (a) et deux MMCs binaires (b)

6 Conclusion

Dans cet article, nous avons présenté un système générique pour extraire de l'information à partir de bandes sons de video, axé sur un décodage en deux étapes et susceptible de détecter des événements sonores simultanés. L'étape de segmentation a été réalisée avec un algorithme basé sur un CIB, et l'étape de classification en utilisant une approche MAP. Nous avons montré que ce système conduit à de bons résultats par comparaison avec le meilleur système testé dans un précédent article et utilisant un algorithme de Viterbi. Nous avons également montré qu'on peut l'améliorer en utilisant les modèles de classe plutôt qu'un algorithme CIB "aveugle" pour la segmentation. Pour illustrer ceci, une segmentation/classification de type Viterbi avec des MMCs à deux états a été proposée. Les résultats étaient comparables aux tests d'hypothèse binaires, malgré les limitations de segmentation propres au Viterbi. Prochainement, nous travaillerons à l'amélioration de la qualité de l'étape de segmentation, et à l'adaptation aveugle des modèles sur de nouvelles données.

Références

- [1] M. Cettolo. Segmentation, classification and clustering of an Italian broadcast news corpus. *Content-based Multimedia Information Access*, 2000.
- [2] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, et S.J. Young. Segment generation and clustering in the HTK Broadcast news transcription system. *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
- [3] G. Gravier M. Betser et R. Gribonval. Extraction of information from video sound tracks - can we detect simultaneous events? Dans *International Workshop on Content Based Multimedia*, 2003.
- [4] Bowen Zhou et John Hansen. Unsupervised audio stream segmentation and clustering via the Bayesian information criterion. Dans *Intl. Conf. Speech and Language Processing*, pages 714–717, 2000.