Construction d'une signature audio pour l'indexation de documents audiovisuels

J.M. Brück[1] S. Bres[1] D. Pellerin[2]

[1] LIRIS, FRE CNRS 2672, INSA de Lyon, 20, Avenue A. Einstein, 69621 Villeurbanne Cedex

[2] LIS, UMR CNRS 5083, Avenue Félix Viallet, 38031 Grenoble Cedex

{jmbruck,stephane.bres}@insa-lyon.fr, denis.pellerin@lis.inpg.fr

Résumé

Ce travail traite du problème général de la recherche de documents audiovisuels dans de grandes bases. L'objectif est de réaliser une signature audio permettant de retrouver un document dans une base à partir d'un extrait sonore. La signature audio présentée ici a été développée pour être rapidement calculée sur une grande base de signaux audio et pour permettre ensuite une recherche rapide. Elle résume l'enregistrement audio à partir d'une analyse en fréquence et allie compacité, efficacité et robustesse.

Un travail similaire et complémentaire a été conduit par notre laboratoire sur la partie « image » des documents audiovisuels. L'association des deux signatures permet une caractérisation plus précise des documents. Nous ne présentons ici que la partie concernant la construction de la signature audio.

Mots clefs

Signature audiovisuelle, monitoring, indexation, identification, recherche rapide.

1 Introduction

L'intérêt des méthodes d'indexation et de recherche de documents audiovisuels n'est plus à justifier. En effet, l'augmentation sans cesse plus importante du nombre de ces documents rend leur gestion difficile voire impossible sans méthodes appropriées. Ces méthodes utilisent toutes les informations disponibles, comme par exemple la partie « image » ou la partie « audio » du document audiovisuel. L'analyse de la partie audio d'une séquence vidéo permet une caractérisation qui peut dans certains cas donner des résultats très précis sur le contenu de cette vidéo [2] [3]. Cette analyse se fait le plus souvent avec un objectif de caractérisation et d'extraction d'informations. C'est donc alors la segmentation et la catégorisation du son qui sont privilégiées [4]. Les classes les plus couramment utilisées sont : les silences, les dialogues, les bruits ambiants et la musique [5] [6].

Nous nous sommes pour notre part, orientés vers l'extraction d'une *signature* qui permette une caractérisation unique d'un extrait quelconque d'une bande audio. Cette étude a été réalisée parallèlement à celle d'une signature image, construite dans le même esprit. L'objectif final est de cumuler les deux « empreintes » pour qu'elles se complètent et ainsi obtenir une signature globale très forte, pour un document audiovisuel. La signature audio que nous proposons est calculée une fois pour toutes et sur toute la durée des documents d'une base. Elle a été développée pour être rapidement calculable, aussi compacte que possible (en nombre d'octets par seconde) et pour permettre d'identifier le document entier mais aussi tous les extraits qui en auront été tirés.

2 La signature audio

Notre objectif est donc de *caractériser* de *manière unique* et dans *son intégralité* une bande son d'un document. Cette caractérisation doit être la plus compacte possible pour limiter le volume de stockage et permettre une recherche rapide. Celle-ci ne s'appuiera que sur le contenu « signal » des documents et n'aura plus de lien « sémantique » avec le document d'origine : vouloir conserver une information sémantique conduirait très certainement à dépasser de beaucoup ce volume minimum.

2.1 Principe

Nous cherchons à *comprimer* la taille de la signature en ne retenant de la source que ce qui peut le mieux la *caractériser* et la rendre *unique*. L'analyse des fréquences contenues dans le signal sonore est très généralement retenue pour sa caractérisation [1]. L'énergie est le plus souvent concentrée en dessous de la fréquence 3 kHz. C'est dans ce domaine que l'on trouvera le plus d'informations dans le signal.

Ce projet étant mené conjointement avec celui de la signature vidéo, il faut trouver un format de données de

taille équivalente et il est surtout obligatoire, pour pouvoir faire coopérer les deux systèmes, d'avoir la même granularité temporelle : cette granularité temporelle sera l'image. Cela permet, lorsque l'extrait est retrouvé dans un fichier source, de dire à quel endroit il se trouve, à l'image près.

Nous avons choisi de réaliser un banc de 8 filtres dans la partie la plus significative du signal et nous avons opté pour une gamme de fréquences allant de 300 Hz à 2000 Hz. Cette bande correspond à la zone la plus sensible de notre audition. Ainsi, on retrouve en téléphonie cette même bande passante et en principe on peut reconnaître la voix d'une personne ou un morceau de musique sans problème. Ces filtres sont construits par fenêtres glissantes. Pour être sûr de ne pas avoir de discontinuités importantes dans les résultats, nous utilisons en pratique une fenêtre glissante avec recouvrement. Pour rester « synchrone » avec la cadence image, nous avons choisi de faire glisser la fenêtre par bonds de 40 ms, ce qui correspond, pour une fréquence d'échantillonnage de 5kHz, à 200 échantillons du signal. Par ailleurs, la transformée de Fourier est calculée sur 1024 points. Cette valeur est un compromis entre les différentes contraintes qui nous sont imposées : puissance de 2 pour l'utilisation de la FFT, une valeur grande devant la plus grande période des parties actives, une localisation temporelle suffisante pour garder l'aspect évolutif du signal… La longueur de cette fenêtre représente environ 5 images.

Nous analysons donc la bande 300Hz – 2kHz à partir d'un banc de 8 filtres. Nous découpons cette bande en huit parties sur la base d'une échelle logarithmique, en référence à la façon dont nous percevons les signaux sonores. Nous obtenons donc huit intervalles sensés être équi-énergétiques « auditivement » parlant. Les fréquences de coupure sont présentées dans le tableau 1.

300	336	377	424	476	600	756	1200	1905

Tableau 1 - *Fréquences de coupure du banc de 8 filtres utilisés pour la construction de la signature d'un signal sonore.*

Au final, chacun des 8 filtres donne une valeur d'énergie, codée sur 16 bits et cela pour chaque image, donc toutes les 40 ms d'audio (Figure 1).

2.2 Recherche d'une séquence à partir d'un extrait

Pour rechercher un extrait dans la base des documents audio, nous fournissons directement cet extrait requête. Sa signature particulière est calculée et comparée avec les signatures présentes dans la base. La comparaison se fait par différence filtre à filtre des données sur la longueur de l'extrait par glissement. Un indice de dissimilarité locale est calculé par somme des valeurs absolues de ces différences. On obtient alors des courbes de dissimilarité du type de celle présentée sur la figure 2, lors de la comparaison d'une requête avec un signal de la base.

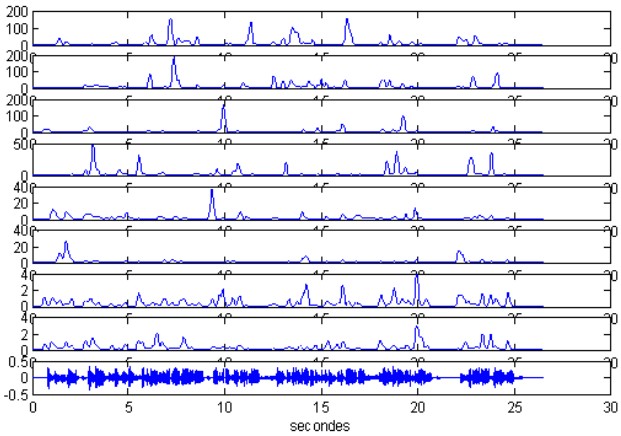

Figure 1 - *Résultat en sortie du banc de 8 filtres pour un signal sonore de 25 secondes dont la représentation temporelle est donnée sur le graphique du bas.*

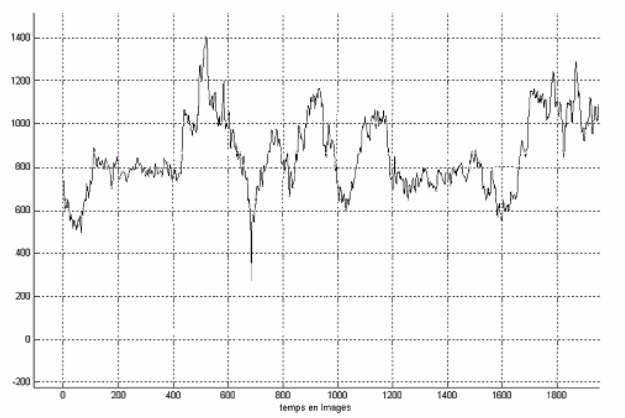

Figure 2 – *Exemple de courbe de dissimilarité entre une requête et un signal audio.*

L'extraction des minimums de dissimilarité localise les positions de correspondance maximum de la requête dans le signal audio testé. Afin d'améliorer l'extraction de ces minimums, nous réalisons un « nettoyage itératif » des données de la courbe de dissimilarité :

- Nous calculons tout d'abord une version lissée en soustrayant, en chaque points de la courbe de dissimilarité, la valeur de dissimilarité à la valeur médiane locale, calculée sur 11 échantillons temporels.
- Nous éliminons les valeurs supérieures à la moyenne de la courbe en les remplaçant par cette valeur moyenne.

Nous appliquons ce processus cinq fois de suite. Les zones de correspondance entre le signal audio requête et la séquence de la base apparaissent alors, de façon nette, par des pics très précis. Un exemple de résultat d'un tel traitement est donné sur la figure 3. Il est obtenu à partir de la courbe de la figure 2. On utilise alors une valeur de seuil pour extraire les positions temporelles des extraits qui correspondent le mieux à la requête dans chaque signal de la base. L'utilisation de cette valeur de seuil permet de ne sélectionner que les correspondances suffisamment pertinentes.

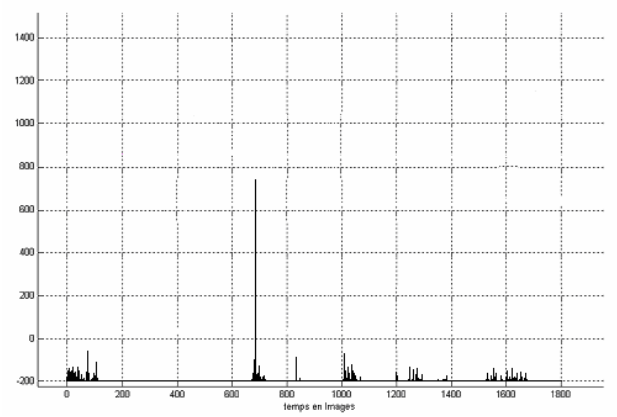

Figure 3 – *Localisation des minimums locaux de la courbe de dissimilarité, par « nettoyage itératif ». Ils correspondent ici aux maxima locaux.*

3 Résultats

Nous avons réalisé les tests sur une partie de notre base qui regroupe plus de 50 Go de documents audiovisuels. Les documents appartiennent à plusieurs catégories de manière à varier le contexte de nos signatures :

- Journaux télévisés
- Publicités
- Documentaires
- Dessins animés
- Séries

Le temps de calcul d'une signature est de l'ordre de 1 minute pour 30 minutes de vidéo traitées. Nous obtenons donc un gain voisin de 30 sur le temps d'extraction de la signature par rapport à la durée de la vidéo elle-même. Nous allons maintenant détailler quelques uns des résultats que nous avons obtenus pour différents types de requêtes.

3.1 Influence de la durée de la requête

Dans le cas d'une sélection de durée trop courte, on se heurte au phénomène de répétition et l'on aura des détections très rapprochées par exemple sur le même motif musical, ou sur des sons répétitifs (Figure 4). Il suffit, pour préciser la requête d'allonger la durée de l'extrait. De ce point de vue, une requête correspondant au son de 200 images (soit une durée de 8 secondes) permet une bonne discrimination, en limitant le nombre de réponses non souhaitées.

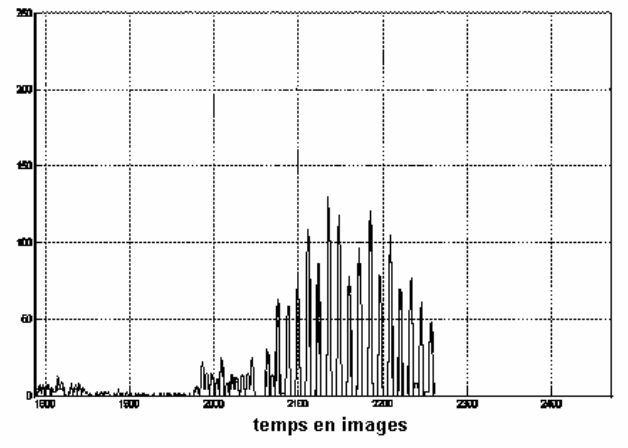

Figure 4 – *La requête est un son répétitif et court (le « ding » d'une sonnette) : beaucoup de candidats pour l'appariement. La requête est trop peu spécifique.*

3.2 Le cas des silences

Lorsque des silences proportionnellement longs sont contenus dans la séquence requête, cela introduit un bruit important dans le résultat de la dissimilarité calculée. Le résultat de la recherche donne tous les passages comportant un silence ! Un silence se caractérise par un motif plat de ses composantes énergétiques. C'est une « signature », mais elle ne caractérise pas de manière unique la source, ce qui implique de très nombreuses détections. On peut contourner ce problème lors du choix du type d'extrait et de sa durée. Mais attention, le silence lui-même, peut aussi caractériser une séquence, donc la sélection a un rôle important dans les résultats que nous obtenons.

3.3 Deux exemples concrets de recherche

Nous avons testé notre signature dans des cas concrets de recherche à partir de requêtes correctement choisies. Nous n'avons pas directement de « vérité terrain » pour notre base de 50 Go de vidéos, qui pourrait nous donner exactement le nombre d'occurrence de chaque requête possible. Cela nous semble par ailleurs très difficile à construire dans le futur, étant donnée la quantité importante d'heures de vidéos que cela représente. Il nous est donc difficile de présenter des résultats sous la forme de courbes Rappel / Précision comme c'est généralement l'usage dans un contexte d'indexation. Cependant, nous avons essayé de faire des tests à partir de requêtes que

nous savions présentes à de nombreuses reprises dans la base pour valider l'efficacité et la robustesse de notre méthode. Nous nous sommes intéressés en particulier aux génériques de séries télévisées dont nous possédions plusieurs épisodes.

La musique du générique de « Friends »
La requête est cette fois-ci un extrait sonore de durée suffisante et ne contient pas de silence. Il s'agit de la musique du générique de la série « Friends », sélectionnée sur plus de 300 images (12 secondes). La recherche est réalisée sur notre base de 50 Go, qui contient 18 épisodes de cette série. Ces 18 fichiers sont les seuls à répondre positivement à la requête. Pour 17 d'entre eux, nous n'avons qu'une seule détection, au moment du générique. Pour le 18^{ième} fichier, à la bonne détection s'ajoute deux autres détections. La première correspond à une reprise de la musique du générique. Il est donc correct que la détection soit faite à nouveau. Pour la seconde en revanche, il s'agit d'un passage musical qui n'est pas complètement identique à la requête, et qui peut être considéré comme une fausse détection.

La musique du générique de « Un gars, une fille »
La musique du générique est prise ici sur 150 images (6 secondes). C'est un générique plus court, et le montage de la série est plus « haché » au départ. Nous obtenons tout de même de bons résultats pour la recherche. Les 11 épisodes que nous avions ont bien été sélectionnés lors de la recherche : 6 d'entre eux ne présentent qu'une seule détection. Pour les 5 autres, nous obtenons respectivement 3, 4, 5, 6 et 11 détections, dont celle correspondant à la requête qui obtient l'amplitude maximale. Les autres extraits sélectionnés contiennent aux moins des bribes de la musique du générique.

4 Conclusion et perspectives

Nous avons cherché à caractériser un signal sonore par une signature à la fois compacte, robuste et unique. Dans ce contexte, c'est bien le signal sonore dans son ensemble que nous cherchons à identifier. Nous n'avons donc pas tenté d'opérer la séparation des différentes sources sonores présentes dans un même extrait. Il n'est donc pas possible de trouver une source donnée, si celle-ci est masquée par d'autres sources. Avec les limitations qu'imposent les hypothèses de départ, nous avons obtenu une signature qui permet une recherche à la fois rapide et précise, et d'autant plus précise que la séquence audio de requête est longue. Les tests que nous avons réalisés nous ont donné des résultats très encourageants, et ont montré en particulier la robustesse de cette signature vis-à-vis de la variabilité qui peut exister entre des signaux sonores identiques mais provenant d'enregistrements différents (comme les génériques de séries télévisées).

A l'heure actuelle, la principale lacune de cette signature réside dans le fait que la recherche dans la base des signatures se fait de façon exhaustive. Cela est dû à l'absence de classement de ces signatures. Il nous semble donc intéressant d'étendre l'utilisation de cette signature à une catégorisation de l'extrait sonore selon différentes classes très générales : Parole / Musique / Bruit / Silence. Cette connaissance servira tout à la fois à l'amélioration de la recherche mais aussi à apporter une information supplémentaire sur la vidéo elle-même, information directement utilisable pour son indexation.
Nous travaillons à l'heure actuelle sur l'évaluation à une plus grande échelle de notre méthode de signature audio, par multiplication des requêtes sur notre base de 50 Go. De plus, nous testons la meilleure façon de faire collaborer cette signature audio avec la signature vidéo développée par ailleurs. L'objectif final est d'aboutir à une indexation réellement multimodale.

Références

[1] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. "A review of algorithms for audio fingerprinting", International Workshop on Multimedia Signal Processing (MMSP'02), St. Thomas, US Virgin Islands, December 2002.

[2] J. Foote, M. Cooper, and L. Wilcox. "Enhanced Video Browsing Using Automatically Extracted Audio Excerpts", Proc. IEEE Int. Conf. on Multimedia and Expo (ICME'02), **2**:357-60, Lausanne, Switzerland, August 2002.

[3] J. Haitsma, T. Kalker, and J. Oostveen, "Robust Audio Hashing for Content Identification", Proc. of the Content-Based Multimedia Indexing (CBMI'01), Rennes, France, September 2001.

[4] F. Kurth, "A Ranking Technique for fast Audio Identification", International Workshop on Multimedia Signal Processing (MMSP'02), St. Thomas, US Virgin Islands, December 2002.

[5] D. Li, I. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval", Pattern Recognition Letters, 22(5):533-544, 2001.

[6] T. Zhang and C.C. Kuo, "Hierarchical classification of audio data for archiving and retrieval", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99), vol. 6, 3001-3004, Phoenix, USA, 1999.