# Identification de signaux audio par appariement de chaînes

Jérome Lebossé[1], Luc Brun[2], Jean Claude Pailles[3]

(1, 2)GREYC UMR 6072, ENSICAEN, 6 Bd du Maréchal Juin,14050 Caen
(1, 3) France Télécom R&D, 42 rue des Coutures,BP 6243 14066 CAEN Cedex 4
(2)luc.brun@greyc.ensicaen.fr,(1, 3){jerome.lebosse,jeanclaude.paille}@orange-ftgroup.com

**Thèmes choisis :** (4.4) Reconnaissance des formes, (4.6) Indexation
**Problème traité :** Identification d'un fichier audio éventuellement altéré dans une base de données de fichiers originaux
**Originalité du travail :** Nous proposons une méthode d'identification basée à la fois sur une découpe adaptative du signal et sur un traitement des erreurs de segmentation à l'aide d'une fonction de similarité entre chaînes.
**Résultats nouveaux :** La fonction de similarité que nous proposons permet à la fois d'identifier un fichier lorsqu'il est présent et de tester sa présence dans la base.

## 1 Introduction

Une empreinte audio est un court code qui permet de retrouver rapidement un document éventuellement altéré (compression, décalages, ...) dans une base de données. Le document altéré est appelé un dérivé du document original [3]. Notons que deux chansons d'un même auteur ne sont pas co-dérivées. De même, une reprise d'une chanson n'est généralement pas un co-dérivé de l'original. Les méthodes d'identification doivent pouvoir identifier un signal à partir d'un court extrait. Il est donc nécessaire de calculer des valeurs caractéristiques (sous-empreintes) tout au long du signal. Les méthodes de définition d'empreintes sont en général basées sur une décomposition du signal en fenêtres de tailles fixes avec recouvrement. Ce type de méthode [2] est sensible aux décalages du signal induits par la sélection aléatoire de l'échantillon pris pour l'identification. D'un autre côté, la définition d'intervalles à l'aide d'une segmentation du signal de type onsets [1] est peu sensible aux décalages mais ne permet d'assurer la détection d'un nombre suffisant d'intervalles sur un court échantillon (typiquement 5 secondes) pour garantir une identification robuste du signal.

Notre idée est donc de combiner les avantages de ces deux approches en définissant une nouvelle méthode de segmentation qui soit à la fois robuste aux altérations telles que le décalage ou la compression et qui fournisse un nombre suffisant d'intervalles pour identifier le signal à partir d'un court échantillon. Cette méthode repose sur la détection de positions particulières dans le signal temporel. Comme le montre la Figure 2(a), un intervalle d'observation $I_o$ est d'abord défini. Nous considérons alors les sous-intervalles ($I_e$) de $I_o$, chaque intervalle $I_e$ étant pondéré par l'énergie moyenne du signal sur celui-ci. L'intervalle $I_{emax}$ ayant l'énergie la plus élevée est sélectionnée pour le calcul de l'empreinte. Un nouvel intervalle $I_o$ est ensuite défini juste après l'intervalle $I_{emax}$ sélectionné. L'utilisation de $I_o$ permet de garantir un taux minimum de détection d'intervalles tandis que l'intervalle $I_{emax}$ permet de synchroniser deux signaux sur les pics significatifs de ceux-ci. Plus de détails peuvent être trouvés dans [5]. Les altérations modifiant sensiblement le signal à l'intérieur de chaque intervalle, nous avons décidé de définir nos sous-empreintes comme l'espace temporel (en ms) entre deux intervalles consécutifs. L'empreinte d'un signal est alors défini comme la suite des distances entre les intervalles ainsi détectés.

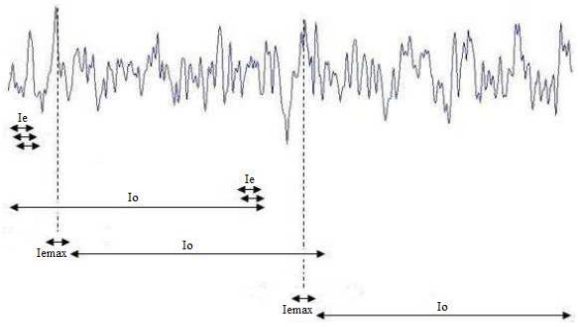


Fig. 1 – Méthode de segmentation audio

## 2 Reconnaissance d'empreintes

Notre méthode d'identification se doit d'être robuste aux insertions/suppressions/modifications de valeurs induites par les altérations du signal. Les méthodes d'appariement de chaînes permettent de mesurer une similarité entre chaînes à partir d'opérations d'insertion, suppression et de substitution. Ces méthodes semblent donc être appropriées. Cependant, une fonction «classique» de score entre chaînes, ne permet pas de différencier une longue séquence de correspondance entre symboles suivie d'une séquence de non correspondance d'une suite alternant correspondances et non correspondances.

Or, dans notre cadre, deux empreintes dérivées d'un même contenu partagent en commun de longues séquences alors que cette distribution est aléatoire pour deux documents indépendants. Nous avons donc défini une fonction de score croissant de manière non linéaire lors de séquences de correspondance afin de favoriser de telles séquences :

$$S(i,j) = \begin{cases} \alpha S(i-1,j-1) + \beta & \text{Si } s_i = s_j \\ \frac{1}{\gamma} max \begin{pmatrix} 0, \\ S(i,j-1), \\ S(i-1,j) \end{pmatrix} - \beta & \text{sinon} \end{cases} \quad (1)$$

Les constantes $\alpha$, $\beta$, $\gamma$ sont déterminées expérimentalement et satisfont la condition $1 < \gamma < \alpha$ afin d'avoir une décroissance du score moins importante que la croissance. $s_i$ et $s_j$ correspondent aux symboles d'index $i$ et $j$ dans les deux chaînes à apparier.

La comparaison d'une empreinte d'entrée avec la base de données peut être effectuée par des méthodes d'alignement de chaînes [3] basées sur notre fonction de score (équation 1). Ce type d'algorithme est généralement accéléré en utilisant une méthode de filtrage basée sur les q-grams [4] (mots de longueur $q$). Toutefois, le théorème de Jokinen-Ukkonen [4] sur lequel s'appuie ce type filtrage s'adapte mal à notre fonction de score qui induit une mémoire des appariements passés. Nous reprenons donc l'idée de base du filtrage par $q$ grams en pondérant chaque $q$ gram par un score défini à partir de l'équation 1. Plus formellement, considérons $Q_{D,I}$ un q-grams entre une empreinte d'entrée I et une empreinte D contenue dans la base de données. Notons également $(I_i)_{i\in\{1,\ldots,p\}}$ et $(D_j)_{j\in\{1,\ldots,q\}}$ les indices dans $I$ et $D$ où apparaît $Q_{D,I}$. La pondération de $Q_{D,I}$ est alors définie en considérant des sous chaînes de longueur $m > q$ par l'équation suivante :

$$score(Q_{D,I}) = \sum_{i=1}^{p} \sum_{j=1}^{q} S(I[I_i, I_i + m - 1], D[D_j, D_j + m - 1]) \quad (2)$$

ou $S(I[I_i, I_i + m - 1], D[D_j, D_j + m - 1])$ représente notre fonction de score (équation 1) calculée entre les deux sous chaînes de $I$ et $D$ de longueur $m$ et commençant respectivement aux indices $I_i$ et $D_j$. Le $q$ gram $Q_{D,I}$ est donc un préfixe de ces deux sous chaînes. La valeur de $m$ choisie dans nos expériences est égale à 20 ce qui correspond approximativement à 1 seconde de signal du fait de notre taux moyen de détection d'intervalles.

Notons à présent, $\{Q_{D,I} \subset D\}$ l'ensemble de q-grams entre $D$ et $I$. Le score de $D$ est défini à partir de la somme des scores des q-grams communs à $D$ et $I$ :

$$score(I, D) = \sum_{Q_{D,I} \subset D} score(Q_{D,I}) \quad (3)$$

Nos expérimentations (Section 3) montrent que l'empreinte de score maximal correspond toujours au co-dérivé de l'empreinte d'entrée lorsque celle-ci est stockée dans la base. Cependant, une méthode d'identification doit être capable de vérifier *la présence* d'un co-dérivé dans la base. Par conséquent, le plus bas score obtenu par un contenu co-dérivé doit être supérieur au meilleur score d'un contenu non dérivé. Cependant, comme le montre nos expériences, le score défini par l'équation 3 ne satisfait pas cette contrainte. Ceci est essentiellement due à la faible valeur de $m$ choisie pour le filtrage. Nous avons donc décidé de considérer uniquement l'empreinte de score maximal retournée par notre étape de filtrage. Nous vérifions ensuite sa similarité avec l'empreinte d'entrée sur une plus longue chaîne ($M > m$) à partir des positions ayant obtenu le meilleur score lors de l'étape de filtrage. Plus formellement, soit $D$ la meilleure empreinte retournée par notre étape de filtrage, et $i_{max}, j_{max}$ deux indices dans $I$ et $D$ tels que $I[i_{max}, i_{max} + q - 1] = D[j_{max}, j_{max} + q - 1]$ et $S(I[i_{max}, i_{max} + m], D[j_{max}, j_{max} + m])$ est maximal pour toutes les positions de $q$ grams communs entre $I$ et $D$. Le score final est alors défini par :

$$score(I) = S(I[i_{max}, i_{max} + M], D[j_{max}, j_{max} + M]) \quad (4)$$

La valeur de $M$ choisie dans nos expériences est égale à 100 ce qui correspond approximativement à 5 secondes de signal.

## 3 Expérimentations

Notre base de données contient plus de 350 chansons d'environ 4 minutes chacune codées à $750Kbps$. Du fait de nos choix pour $I_0$ ($100ms$), $I_e(1ms)$ (section 1) et du pas d'échantillonnage choisi pour mesurer les distances entre intervalles

| Kbps | 48 | 64 | 96 | 128 | 192 | 256 |
|---|---|---|---|---|---|---|
| Notre méthode | 80 | 83 | 85 | 87 | 89 | 94 |
| Haitsma | 5 | 7 | 12 | 24 | 25 | 30 |

(a) Performances des méthodes d'extraction d'empreintes

| | Classements de filtrage | | | Scores finaux | | |
|---|---|---|---|---|---|---|
| Scores | $1^{ier}$ | $2^{nd}$ | $3^{ieme}$ | $1^{ier}$ | $2^{nd}$ | $3^{ieme}$ |
| Min | **14878** | 0 | 0 | **100000** | 0 | 0 |
| Moy | $11.10^5$ | 240.72 | 143.74 | $3.10^{15}$ | 16 | 3 |
| Max | $4.10^6$ | $\mathbf{20.}10^3$ | $10.10^3$ | $5.10^{17}$ | **2800** | 590 |

(b) scores d'identification

($\frac{1}{44100}$) chaque sous empreinte nécessite 13 bits pour être codée. De plus, le taux moyen de détection d'intervalles étant égal à 21 intervalles par secondes, la taille de l'empreinte correspondant à une minute de signal est égale à $2,1Ko$. Pour évaluer les performances de notre algorithme d'extraction d'empreinte, nous avons compressé chaque fichier audio à différents taux (48, ..., 256 Kbps). Les performances d'une méthode d'extraction d'empreintes sont alors mesurées par le *taux de reconnaissance* défini comme le pourcentage de sous-empreintes communes entre deux empreintes issues de contenus co-dérivés.

Le tableau 2(a) représente le taux de reconnaissance obtenu par notre méthode comparé à celle de Haitsma. Comme le montre ce tableau, le taux le plus élevé obtenu par la méthode d'Haitsma (30% à $256Kbps$) est inférieur au taux le plus bas obtenu par notre méthode (80% à $48Kbps$). La robustesse à la compression de notre méthode peut être en partie expliquée par le fait que la définition des sous valeurs basée sur l'écart temporel entre intervalles détectés est moins sensible aux altérations qu'une définition basée sur le contenu du signal dans chaque intervalle. D'autres évaluations de notre méthode d'extraction concernant notamment la robustesse au décalage peuvent être trouvées dans [5].

### 3.1 Reconnaissance d'empreinte

Pour ces expériences, 10 secondes de signal audio ont été extraites aléatoirement de notre base de données puis compressées à 128Kbps. L'empreinte de ces 10 secondes a été calculée pour chaque signal. Notre méthode de reconnaissance (Section 2) a été mise en œuvre pour identifier chaque empreinte d'entrée. La première moitié de chaque empreinte (correspondant à 5 secondes) est utilisée par l'étape de filtrage. La seconde partie permet de compléter l'empreinte d'entré dans le cas où $i_{max}$ n'est pas égal à 0 lors de l'évaluation de l'équation 4. Notons que pour ces tests, chaque empreinte a un co-dérivé dans la base de données correspondant à son original non compressé. La taille minimale d'un $q$-grams pour notre procédure de filtrage a été fixée à 5. Les valeurs de $\alpha$,$\gamma$ et $\beta$ ont été respectivement mises à : $\alpha = 1.5$ ; $\gamma = 1.1$ ; $\beta = 20$.

Les trois premières colonnes de la Table 2(b) montrent les valeurs minimales, moyennes et maximales des scores de filtrage (équation 3) obtenus par les trois empreintes de la base de données de score maximal en fonction de l'équation 3. Dans cette expérience, l'empreinte de la base de données classée première a toujours correspondu au co-dérivé de l'original. Le score de la seconde empreinte correspond donc au meilleur score qui serait obtenu si une version co-dérivée de l'entrée n'était pas présente dans la base. Ce type d'empreinte est appelé un *meilleur faux positif*. Comme le montre la seconde ligne de la table, le score du co-dérivé est généralement plus élevé que ceux obtenus par les autres empreintes. Cependant, le score obtenu par le meilleur faux positif pour certaines entrées peut être supérieur à celui obtenu par le co-dérivé pour d'autres entrées (cellules (Min, $1^{ier}$) et (Max, $2^{nd}$) Table 2(b)). Ce dernier point ne permet donc pas de vérifier la présence d'un co-dérivé dans la base de données.

Les trois dernières colonnes, de la Table 2(b) montrent les valeurs minimales, moyennes et maximales des scores calculés en fonction de l'équation 4 sur les trois empreintes retournées par notre étape de filtrage. Rappelons que le but de cette équation est d'augmenter l'écart entre les scores du co-dérivé et des autres empreintes. La réalisation de cet objectif est confirmée par la seconde ligne de la table 2(b) qui montre un accroissement de l'écart entre le score de l'empreinte co-dérivée et celle du meilleur faux positif. De plus comme le montre les cellules (Min, $1^{ier}$) et (Max, $2^{nd}$) le meilleur score du meilleur faux positif est inférieur au plus faible score d'une empreinte co-dérivée. Un seuil défini entre ces deux scores permet donc de vérifier la présence d'un document audio co-dérivé dans la base de données.

## Références


[1] S. Hainsworth and M. Macleod. Onset detection in musical audio signals. In *Proc. of the International Computer Music Conference*, 2003.

[2] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *Proc. of the International Sympyposium on Music Information Retrieval*, pages 144–148, 2002.

[3] T. Hoad and J. Zobel. Video similarity detection for digital rights management. In *Proc. of the Australasian Computer Science Conference*, pages 237–245, 2003.

[4] P. Jokinen and E. Ukkonen. Two algorithms for approximate string matching in static texts. In *Lecture Notes in Computer Science*, pages 240–248, 1991.

[5] J. Lebosse, L. Brun, and J. Pailles. A robust audio fingerprint extraction algorithm. In *Proc. of the Conf in SPPRA*, Innsbruck(Austria), February 2007.