Identification de signaux audio par appariement de chaînes

Jérome Lebossé[1], Luc Brun[2], Jean Claude Paillès[1]

[1]France Télécom R&D
42 rue des Coutures, BP 6243, 14066 CAEN Cedex 4, France

[2]GREYC UMR 6072, ENSICAEN
6 Boulevard du Maréchal Juin, 14050 Caen, France

(1){jerome.lebosse,jeanclaude.pailles}@orange-ftgroup.com
(2)luc.brun@greyc.ensicaen.fr

Résumé – Le fingerprint audio est un court résumé d'un document audio calculé à partir des propriétés du signal. Comme l'empreinte digitale humaine, le fingerprint audio permet d'identifier un document audio parmi un lot de candidats sans en déduire aucune autre caractéristique. Dans cet article, nous proposons une méthode d'extraction de fingerprint basée sur une nouvelle méthode de segmentation adaptative du signal. La combinaison d'une méthode d'appariement de chaîne avec un pré-filtrage par q-grams permet d'identifier un extrait audio inconnu et de décider si cet extrait est une version dérivée d'un fingerprint préalablement calculé et stocké ou si aucun fingerprint de la base de donnée de correspond à l'extrait d'entrée.

Abstract – An Audio fingerprint is a small digest of an audio file computed from its main perceptual properties. Like human fingerprints, audio fingerprints allow to identify an audio file among a set of candidates without retrieving any other characteristics. We propose in this paper a fingerprint extraction algorithm based on a new audio segmentation method. The identification task is performed by using string matching algorithms combined with a q-grams filtration to decide if an input signal is a derived version of a stored fingerprint or if no database fingerprint corresponds to the input signal.

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 [4]. 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.

Ce genre de méthode peut être utilisé pour diverses applications comme l'archivage, la surveillance réseau, ou la gestion des droits d'auteurs où nous nous sommes investis. Notre idée est de stocker automatiquement, au sein de l'ordinateur personnel, les fingerprints de chaque morceau de musique acquis légalement. Ensuite, lorsque l'utilisateur souhaite écouter un document audio compressé ou gravé sur un CD, le fingerprint est calculé en parallèle à la lecture et comparé avec ceux stockés. Si le fingerprint est identifié, la lecture continue. Sinon, l'écoute est stoppée. Cette application implique donc une très faible sensibilité du fingerprint envers tous les types de compression (mp3, ogg, wma, ...). Le fingerprint doit aussi être identifié à partir d'extraits de 5s pris à n'importe quel moment de la lecture, impliquant une robustesse aux décalages temporels. De plus, la taille du fingerprint et la rapidité de calcul et de reconnaissance sont des critères déterminants pour une utilisation sur ordinateurs familiaux ou mobiles.

Dans cet article, nous commencerons par rappeler notre méthode de calcul de fingerprint en Section 2 et exposerons ensuite notre nouvelle méthode de reconnaissance d'extraits musicaux. La robustesse et l'efficacité de cette méthode seront démontrés à travers les expérimentations de la Section 4.

2 Définition du fingerprint

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 [3] 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 [2] 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 :.

Cette technique de segmentation peut être décomposée en 3 étapes (Figure 1) :

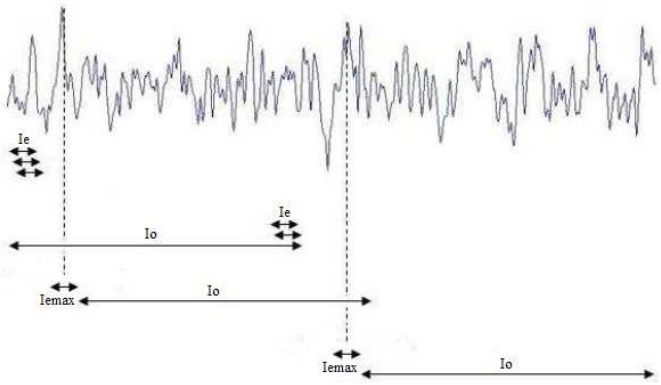

FIG. 1 – Méthode de segmentation audio

- Dans la première étape, un intervalle appelé, Intervalle d'Observation, (I_o) est sélectionné au début de l'extrait. La longueur de cet intervalle est typiquement égale à quelques centaines de millisecondes.
- Ensuite, l'allure du signal de l'intervalle I_o est analysée et divisée en intervalles plus courts (de quelques millisecondes) se recouvrant totalement, appelés Intervalles d'Énergie I_e. L'énergie de chaque intervalle I_e est calculée en prenant la moyenne des échantillons contenus dans l'intervalle..
- Enfin, l'intervalle I_e ayant l'énergie maximale (I_{emax}) est extrait et le prochaine intervalle I_o est sélectionné à la suite du I_{emax}.

Cette méthode permet alors de rapidement synchroniser le processus d'extraction de deux fingerprints de contenus co-dérivés et, par conséquent, d'être robuste aux altérations temporelles (suppressions ou ajout de parties, décalages). L'utilisation de I_o permet alors de garantir un taux minimum de détection d'intervalles tandis que l'intervalle I_{emax} permet de synchroniser synchroniser le processus d'extraction de deux fingerprints de contenus codérivés sur les pics significatifs de ceux-ci et, par conséquent, d'être robuste aux altérations temporelles (suppressions ou ajout de parties, décalages). Plus de détails peuvent être trouvés dans [6].

Les méthodes existantes pour définir les valeurs de sous-fingerprint pour chaque intervalles se basent souvent sur une approximation des fréquences ([3],[1]. Cependant, pour être insensibles aux 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. Si le nombre d'intervalles détectés est égal à n, l'empreinte d'un signal est alors définie comme la suite des distances $s_1 ... s_{n-1}$, calculée en millisecondes, qui sépare deux intervalles successifs.

3 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 α, β, γ 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.

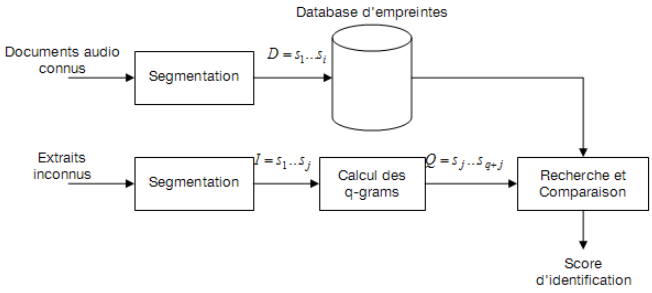

FIG. 2 – Identification d'un fingerprint audio

La comparaison d'une empreinte d'entrée avec la base de données peut être effectuée par des méthodes d'alignement de chaînes [4] 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 [5] (mots de longueur q). Toutefois, le théorème de Jokinen-Ukkonen [5] 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-gram entre une empreinte d'entrée I et une empreinte D contenue dans la base de données. Notons également $(I_i)_{i\in\{1,...,p\}}$ et $(D_j)_{j\in\{1,...,q\}}$ 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 4) 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.

4 Expérimentations

Notre base de données contient plus de 350 chansons d'environ 4 minutes chacune codées à $750Kbps$. Les chansons ont été choisies pour couvrir tous les styles musicaux, du classique au rock, et certaines d'entre elles sont en version originale ainsi qu'en version live. 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 ($\frac{1}{44100}$) 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 fingerprints identiques* défini comme le pourcentage de sous-empreintes communes entre deux empreintes issues de contenus co-dérivés.

TAB. 1 – Performances des méthodes d'extraction d'empreintes

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

Le tableau 1 représente le taux de fingerprint identiques entre les fingerprints de deux contenus co-dérivés 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. Ces résultats doivent toutefois être relativisés en prenant en considération le fait que la méthode de reconnaissance d'empreinte utilisée par Haitsma est basée sur une distance de Hamming calculée à partir de deux symboles communs à deux fingerprints. La méthode de Hamming étant peu sensibles à un des inversions de bits, son utilisation compense la méthode d'extraction de fingerprint. D'autres évaluations de notre méthode d'extraction concernant notamment la robustesse au décalage peuvent être trouvées dans [6].

4.1 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 3) a été mise en oeuvre 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 α,γ et β ont été respectivement mises à : $\alpha = 1.5$; $\gamma = 1.1$; $\beta = 20$. Ces valeurs ne sont pas impératives. Le rôle de β est d'initialiser la croissance du score, il doit être simplement être supérieur à 1. Ensuite, α doit être supérieur à γ afin de favoriser les grandes suites de valeurs communes.

Les trois premières colonnes de la Table 2 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'équa-

TAB. 2 – scores d'identification

	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	**20.**10^3	10.10^3	5.10^{17}	**2800**	590

tion 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). 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 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 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. Notons aussi que lors de l'identification d'un extrait d'une version originale d'une chanson, la version live de cette même chanson obtient un faible score et ces deux oeuvres sont donc bien considérées comme différentes.

5 Conclusions

Dans cet article, nous avons présenté une nouvelle méthode d'identification de signature audio basée sur un algorithme de segmentation du signal associé à une nouvelle fonction de calcul de score de similarité. L'algorithme de segmentation retrouve des instants significatifs dans le signal et assure par la même occasion un taux relativement constant de détection de ces instants. Chaque sous-fingerprint est alors défini par le temps (en ms) de signal écoulé entre deux instants successifs détectés. Ensuite, une première étape de filtrage utilisant les q-grams ainsi qu'une seconde étape d'appariement de chaînes permet de retrouver le fingerprint de la base de données de plus proche de celui du signal d'entrée. Un dernier score basé sur une nouvelle fonction d'évaluation du degré de similarité permet de vérifier si le signal d'entrée est bien une version co-dérivée de l'élément le plus proche appartenant à la base de données. Dans nos prochains travaux, nous projetons d'améliorer la méthode d'indexation de notre base de données en concordance avec l'utilisation des q-grams utilisés dans l'étape de filtrage.

Références

[1] J. Platt C. Burges and S. Jana, "Distorsion discriminant analysis for audio fingerprnting," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 3, pp. 165-174, 2003.

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

[3] 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.

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

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

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