Une empreinte audio à base de CQT appliquée à la surveillance de flux radiophoniques

Sébastien FENET, Yves GRENIER, Gaël RICHARD

Institut TELECOM, TELECOM ParisTech, CNRS-LTCI
37 rue Dareau, 75014 Paris, France
Sebastien.Fenet@Telecom-ParisTech.fr, Yves.Grenier@Telecom-Paristech.fr
Gael.Richard@Telecom-Paristech.fr

Résumé – L'extraction d'empreinte audio s'inscrit dans la problématique plus large de l'identification audio, qui consiste à retrouver des méta données à partir d'un extrait audio. Dans cet article, nous présentons notre algorithme d'extraction d'empreinte audio, appliqué à la détection d'évènements référencés dans un flux. Tout en s'appuyant sur la méthode de Wang [8], notre approche présente une robustesse accrue au glissement fréquentiel grâce notamment à l'utilisation d'une "transformée à Q constant" [2]. Nous montrons finalement que le changement de représentation proposé permet de bien meilleurs taux de détection dans un cas concret de surveillance de flux radiophonique.

Abstract – Audio fingerprint lies in the field of audio identification. It consists in retrieving meta data associated with an unknown audio excerpt. In this article, we present our audio fingerprint algorithm applied to a broadcast monitoring use case. The method we present is inspired by Wang's work [8], in which we have introduced the use of a "constant Q transform" [2]. This allows us to increase the robustness to pitch-shifting. Finally, we show that the proposed fingerprint leads to a better detection score in a broadcast monitoring oriented evaluation.

1 Introduction

L'extraction d'empreinte audio s'inscrit dans la problématique plus large de l'identification audio, qui consiste à retrouver des méta données (artiste, nom de l'album, nom de la chanson, etc.) à partir d'un extrait audio. L'identification audio a reçu une attention accrue de la communauté scientifique ces dernières années du fait de son rôle clé dans de nombreuses applications dont les plus populaires sont : l'identification d'un extrait audio capturé par téléphone mobile [8] et la surveillance automatique de contenu multimédia protégé.

Les méthodes à extraction d'empreinte ramènent le problème de l'identification audio à celui d'une comparaison d'empreintes. Le système dispose d'une base de références audio. Il extrait de chacune de ces références une empreinte, constituant ainsi une base d'empreintes. Lors de l'identification d'un extrait inconnu, le système calcule son empreinte puis cherche l'empreinte la plus proche dans la base. Les enjeux de cette approche sont : la création d'une empreinte qui assure une bonne reconnaissance des extraits audio malgré la présence de distorsions induites par le canal de transmission [4] et l'intégration de cette empreinte dans un système de recherche puissant. Il s'agit en effet d'être capable d'identifier en temps réel une référence audio dans une base de données qui en contient plusieurs centaines de milliers.

L'état de l'art en identification audio par extraction d'empreinte est relativement fourni (voir [3]). On trouve ainsi dans la littérature une palette de systèmes aux approches très diverses, des plus anciens proposant des empreintes relativement légères, basées par exemple sur des courbes de puissance du signal [7], jusqu'à des systèmes extrêmement complexes proposant d'indexer le signal en utilisant de nombreux descripteurs audio (mélodie principale, rythme, timbres, ...)[1], pour arriver à des systèmes moins exhaustifs mais plus rapides donnant lieu à des implémentations industrielles, tels que ceux proposés par Wang [8] ou Haitsma [6].

Dans cet article, nous présentons notre algorithme de détection d'évènements référencés dans un flux audio [1]. Tout en s'appuyant sur la méthode de Wang [8], notre approche permet d'atteindre une robustesse accrue au glissement fréquentiel, distorsion largement présente dans les diffusions radiophoniques. Nous proposons également l'ajout d'un module de post-traitement permettant de discriminer très efficacement les fausses alarmes. Le document s'articule autour du plan suivant : nous détaillons dans un premier temps le cas d'utilisation « surveillance de flux audio » avec ses principes et ses contraintes, nous explicitons ensuite le système que nous avons mis en place, puis nous donnons ses résultats dans le cas d'une surveillance de flux radiophonique.

2 Surveillance de flux

Nous nous plaçons ici dans le cadre d'une surveillance de flux audio. La tâche consiste pour le système à extraction d'empreinte à détecter au sein d'un flux continu toutes les diffu-

1. Ces travaux ont été réalisés dans le cadre du programme QUAERO, financé par OSEO, agence française pour l'innovation.

sions des références dont il dispose dans sa base. Plus concrètement, nous nous attacherons à détecter un ensemble de musiques d'une base de données au sein du flux d'une station de radio. Notons que le flux sonore est constitué de plages temporelles contenant les références à détecter, mais également de plages temporelles où des éléments non référencés sont diffusés (parole, publicités, musiques non référencées). Si nous désignons par $m_1, ..., m_N$ les références de la base, et par b l'ensemble des autres diffusions (considérées comme du bruit du point de vue de l'algorithme), la tâche peut se représenter comme en figure 1.

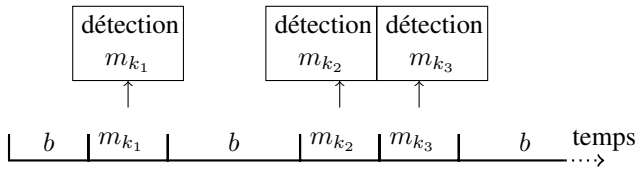

FIGURE 1 – Surveillance de flux radiophonique

Il faut noter que les stations de radio appliquent une série de traitements aux pistes audio qu'elles diffusent [4], tels que la conversion de formats, l'égalisation, la compression d'amplitude, la modification de stéréo ou le rehaussement fréquentiel. La très grande majorité des méthodes de l'état de l'art en identification audio par empreinte est robuste aux distorsions précédentes. Cependant, la robustesse au glissement fréquentiel est nettement moins systématique [5]. Les stations de radio changent fréquemment la vitesse de lecture des morceaux de musique en fonction des contraintes horaires de l'émission qui les diffuse. La technique utilisée pour mettre en œuvre ce changement de vitesse provoque généralement un effet de bord : le glissement fréquentiel. Toutes les fréquences du morceau sont multipliées par un facteur constant. Nous nous attachons particulièrement à proposer une méthode robuste au glissement fréquentiel.

3 Système

3.1 Architecture générale

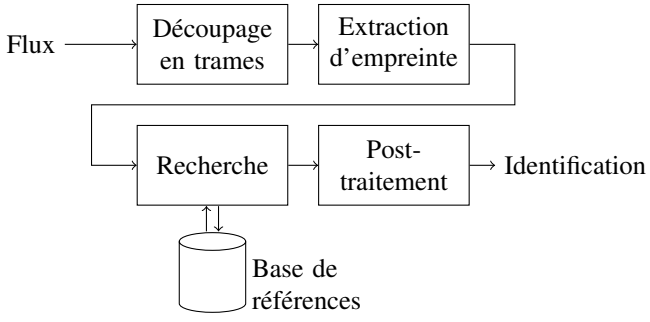

FIGURE 2 – Architecture générale de l'algorithme

Le système doit traiter des flux continus. Ainsi, une première étape (voir figure 2) consiste à découper le signal en trames (appelées dans la suite trames d'analyse) de longueur l_a avec un recouvrement o_a. Ces trames sont traitées par le module *extraction d'empreinte*. Le module de *recherche* se charge ensuite de trouver la référence dans la base avec l'empreinte la plus proche de celle de l'extrait inconnu. Un module de *post-traitement* permet, via la fusion d'identifications successives, de discriminer les requêtes hors-base (cas d'un extrait à identifier qui n'est pas référencé dans la base).

3.2 Extraction d'empreinte

Le schéma retenu pour le calcul de l'empreinte s'appuie sur la méthode exposée par Wang [8] : calculer un spectrogramme, le rendre binaire par une technique d'extraction de pics puis extraire des paires de points à 1 au sein de ce spectrogramme binaire. Toutefois, nous modifions le spectrogramme et la façon d'encoder les paires de points proposés par Wang.

Le spectrogramme que nous utilisons est obtenu par l'application de « transformées à Q constant » [2] successives. Au même titre que la transformée de Fourier, la transformée à Q constant donne l'énergie contenue dans un signal suivant chaque bande fréquentielle. Mais a contrario de la transformée de Fourier qui propose une répartition linéaire des bandes de fréquence, celles de la transformée à Q constant sont réparties logarithmiquement. Ainsi cette représentation est nettement plus adaptée à la description d'évènements acoustiques. Notons en effet que la perception humaine des fréquences suit une échelle proche du logarithme. Notamment, les notes de musique de la gamme occidentale sont espacées géométriquement. De fait, la transformée à Q constant génère un nombre de bandes constant par note. Nous utilisons une transformée à Q constant avec une résolution de 3 bandes fréquentielle par note. Notons également que dans le domaine de la transformée à Q constant, le glissement fréquentiel devient une translation.

Nous extrayons de ce spectrogramme une représentation binaire construite comme suit. Nous pavons le spectrogramme avec des rectangles de largeur ΔT secondes et de hauteur ΔB bandes de fréquences. Dans chacun de ces rectangles, le point d'énergie maximum est mis à 1, le reste des points à 0. Le résultat est un spectrogramme binaire contenant quelques points à 1 qui correspondent aux points d'énergie maximale dans le spectrogramme initial. Notons que cette façon de procéder garantit une densité de points à 1 homogène, tant en fréquence qu'en temps. Cette représentation est donc robuste aux compressions d'amplitude (variation du gain en fonction du temps) et à l'égalisation (variation du gain en fonction la fréquence). De plus, le fait de ne considérer que des points possédant une énergie importante rend la représentation robuste à la plupart des bruits additifs.

3.3 Recherche

3.3.1 Indexation des références

Toutes les références sont préalablement indexées. Utiliser un index permet en effet lors de la phase d'identification d'éviter un parcours exhaustif de la base (prohibitif en termes de temps de calcul dans le cas de très grandes bases). Conformément à la suggestion de Wang [8] nous utilisons des paires de points à 1 du spectrogramme binaire comme clé d'index. L'encodage que nous proposons pour une paire de points est le suivant : soient deux points du spectrogramme binaire avec des dates d'occurrence t_1 et t_2 et situés dans les bandes fréquentielles b_1 et b_2, la paire constituée par ces deux points est représentée par le vecteur

$$[\widehat{b_1}; b_2 - b_1; t_2 - t_1]$$

avec $\widehat{b_1} = \left\lfloor \frac{b_1}{6} \right\rfloor$ une version sous-résolue de b_1.

La première composante ($\widehat{b_1}$) du vecteur proposé situe la paire fréquentiellement, suffisamment approximativement (la résolution initiale est divisée par 6) pour ne pas être sensible à un éventuel glissement fréquentiel. La deuxième composante donne l'étendue fréquentielle de la paire de pics. Notons que cette composante est robuste au glissement fréquentiel. En effet, en cas de glissement fréquentiel, b_1 devient $b_1 + K'$, b_2 devient $b_2 + K'$. Nous avons bien $(b_2 + K') - (b_1 + K') = b_2 - b_1$. La troisième composante donne l'étendue temporelle de la paire.

3.3.2 Identification

La démarche pour identifier un extrait inconnu est la suivante. Le système extrait les paires de pics spectraux de l'extrait inconnu et les associe à leur date d'occurrence dans l'extrait inconnu. A ce stade, le problème est ramené à l'identification d'un ensemble de paires de pics spectraux couplées à leur date d'occurrence : $\{(p, t_{p,inc}),\ p$ paire de pics spectraux de l'extrait inconnu à la date $t_{p,inc}\}$.

Pour chaque paire p, une requête est adressée à l'index. En réponse à cette requête, nous récupérons la liste des références contenant cette paire avec, pour chaque référence, la date d'occurrence de la paire en question dans la référence. Si nous désignons par h la fonction d'index, nous avons :

$$h : p \longmapsto \{(m_i, t_{p,m_i}),\ m_i \text{ possède la paire } p \text{ à la date } t_{p,m_i}\}$$

Si l'extrait inconnu est une version tronquée de la référence m_0 avec un décalage temporel d, alors les paires de l'extrait inconnu se retrouveront statistiquement très nombreuses dans la référence m_0. En ce qui concerne leur date d'occurrence, une paire p ayant pour date d'occurrence $t_{p,inc}$ dans l'extrait inconnu devrait se retrouver dans la référence à la date $t_{p,m_0} = t_{p,inc} + d$. C'est cette propriété que nous utilisons pour réaliser l'identification.

Pour chaque référence m_i, nous stockons les quantités $\{t_{p,m_i} - t_{p,inc},\ p$ paire de l'extrait inconnu$\}$ sous forme d'histogramme (un histogramme par référence). Une fois les histogrammes construits nous cherchons celui qui possède le plus grand maximum. La référence correspondant à cet histogramme est alors considérée comme le meilleur candidat pour identifier l'extrait inconnu. Nous en déduisons également que si l'extrait inconnu lui correspond, la référence a été tronquée avec un décalage correspondant à l'argument du maximum de l'histogramme.

3.4 Post-traitement

Quelque soit l'extrait inconnu à identifier, l'étape précédente renvoie un meilleur candidat. Cela signifie en particulier que c'est le cas même lorsque l'extrait inconnu n'appartient pas à la base. Nous proposons donc d'ajouter une étape de post-traitement destinée à discriminer les extraits qui ne sont pas dans la base. L'approche intuitive consisterait à mettre en place un seuil sur le nombre de paires en commun entre le meilleur candidat et l'extrait inconnu. Dans la pratique il s'avère très difficile voire impossible de fixer un seuil qui fonctionne bien. Dans le cas de données réelles, la variabilité du nombre de paires en commun entre un extrait distordu et la référence correspondante est en effet très importante. Par ailleurs, un tel seuil ne fonctionnerait que dans un cadre de distorsions défini, et il faudrait le retravailler à chaque changement de contexte.

Ainsi nous proposons un mécanisme de post-traitement basé sur la fusion de décisions locales. Nous observons un horizon de D trames d'analyse $\{a_j\}_{j=1..D}$. Chacune d'elle donne un résultat d'identification $(m_j, \Delta t_j)$. Nous organisons alors un vote majoritaire sur l'horizon : si plus de S identifications sont cohérentes (*i.e.* elles pointent vers la même référence, et la quantité $\Delta t_j - j.l_a.(1-o_a)$ est constante), alors l'identification est bonne. Sinon, il s'agit d'un extrait non présent dans la base.

Le réglage de S (seuil de détection) conditionne la sensibilité du système. S peut prendre n'importe quelle valeur entière entre 0 et D. Si $S = D$, nous imposons que toutes les identifications de l'horizon de décision soient cohérentes. Ainsi, nous minimisons très fortement le risque de fausse alarme. En revanche, nous risquons de manquer des détections. A l'inverse, pour S très faible, nous générons beaucoup de fausses alarmes mais nous maximisons les détections. Dans la pratique une valeur efficace pour S est :

$$S = \left\lceil \frac{D}{2} \right\rceil$$

4 Expériences

Nous nous appuyons sur le protocole mis en place dans le cadre de l'évaluation 2010 de OSEO-Quaero. La base de références est constituée de 7309 morceaux de musiques. Le flux de test est celui d'une radio française (RTL) pendant 7 jours, pour lequel nous possédons les annotations de diffusion des morceaux référencés, précisant l'heure de départ et l'heure de fin pour chaque diffusion. La tâche de l'algorithme est de signaler les diffusions de morceaux référencés. Pour chaque détection d'un morceau de la base de références, l'algorithme fournit

l'identifiant du morceau ainsi que la date de détection dans le flux. Si la date de détection est comprise entre le départ annoté et la fin annotée d'une diffusion du même morceau, celle-ci devient une *diffusion détectée* (on comptabilise au plus 1 détection par diffusion d'un morceau de la base). Si l'algorithme détecte un morceau pendant une plage non annotée, ou annotée avec un autre morceau, une fausse alarme est comptabilisée (nous ne limitons pas la comptabilisation des fausses alarmes).

Notons qu'il arrive parfois qu'un titre soit diffusé dans la radio dans une version alors que le même titre est présent dans la base de références dans une autre version (par exemple : le titre est exécuté en direct sur la radio, alors que la base de références contient la version studio). Dans ce cas, nous ne demandons pas à l'algorithme de faire le rapprochement entre les deux versions. En effet, la reconnaissance d'interprétations différentes du même titre est considérée comme hors du périmètre de l'identification audio.

Nous avons comparé deux algorithmes suivant le protocole ci-dessus. Le premier est notre implémentation de la méthode originale de Wang [8] associée au module de tramage et de post-traitement décrits en section 3. Le deuxième est le système décrit dans cet article. Les deux systèmes utilisent les mêmes paramètres pour les modules de tramage et de post-traitement.

Nous utilisons des trames d'analyse de 5s avec un taux de recouvrement de 0.5. Nous prenons $D = 6$ (autrement dit, nous prenons une décision de détection sur 15s de signal) et $S = 3$.

TABLE 1 – Résultats des algorithmes

Algorithme	Diff. détectées / Total	Fausses Alarmes
Original (FFT)	381 / 459 (=83.0%)	0
Proposé (CQT)	447 / 459 (=97.4%)	0

Les résultats (voir table 1) montrent que le taux de détection est bien supérieur avec notre empreinte. D'après notre analyse auditive, cela provient de la présence en quantité non négligeable dans les flux radio de morceaux ayant subi un glissement fréquentiel. En conséquence, ces résultats montrent que notre proposition est robuste non seulement aux mêmes distorsions que l'algorithme original, mais également au glissement fréquentiel. Concernant le post-traitement, nous observons que celui-ci remplit très bien son rôle étant donné qu'il a éliminé toutes les potentielles fausses alarmes tout en conservant un taux de détection très élevé.

5 Conclusion

Dans cet article, nous avons proposé la mise en place de bout en bout d'un système d'identification audio par extraction d'empreinte appliqué à la surveillance de flux. L'empreinte utilisée est inspirée de la méthode de Wang [8] dont nous avons reproduit le schéma d'indexation basé sur des paires de points spectraux. Cependant notre proposition d'utilisation de la transformée à Q constant [2] et l'encodage des paires que nous suggérons permettent d'augmenter significativement la robustesse de la méthode au glissement fréquentiel. Ceci se traduit au niveau des résultats par des scores de détection nettement plus élevés dans un cas d'utilisation type "surveillance de flux radiophonique".

Dans nos expérimentations à venir, nous mettrons l'accent sur l'augmentation de la taille de la base de références afin de prouver la capacité du système à passer à des échelles "industrielles". Quant à nos développements futurs, nous nous concentrerons sur le rapprochement de versions différentes du même titre exposé dans la section 4. Les flux radiophoniques utilisés pour nos expérimentations contiennent en effet 7% de titres interprétés en direct dont l'algorithme possède la version studio dans sa base. Il sera intéressant d'étudier une extension de l'algorithme capable de rapprocher les uns avec les autres, probablement en introduisant dans le modèle des caractéristiques du signal possédant plus de sens sémantique.

Références

[1] T. L. Blum, D. F. Keislar, J. A. Wheaton, and E. H. Wold. Method and article of manufacture for contentbased analysis, storage, retrieval, and segmentation of audio information. US Patent No. 5,918,223, 1999.

[2] J. C. Brown. Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America*, 89(1) :425–434, January 1991.

[3] P. Cano, E. Battle, T. Kalker, and J. Haitsma. A Review of Algorithms for Audio Fingerprinting. *Multimedia Signal Processing, 2002 IEEE Workshop on*, pages 169 – 173, Dec. 2002.

[4] P. Cano, E. Battle, H. Mayer, and H. Neuschmied. Robust Sound Modeling for Song Detection in Broadcast Audio. *Proceedings of AES, 112th Convention, 2002*, page 5531, Munich, Germany, May 2002.

[5] E. Dupraz and G. Richard. Robust frequency-based audio fingerprinting. *ICASSP 2010*, pages 2091–2094, Dallas,USA, March 2010.

[6] J. Haitsma, T. Kalker, and J. Oostveen. Robust audio hashing for content identification. *Content-Based Multimedia Indexing (CBMI)*, Brescia, Italy, September 2001.

[7] J. Lourens. Detection and Logging Advertisements using its Sound. *Communications and Signal Processing, 1990. COMSIG 90. Proceedings., IEEE 1990 South African Symposium on*, pages 209 – 212, Jun 1990.

[8] A. Wang. An Industrial-strength Audio Search Algorithm. *ISMIR 2003, 4th Symposium Conference on Music Information Retrieval*, pages 7 – 13, Baltimore, Maryland, USA, October 2003.