Détection de jingles dans les documents sonores

Julien PINQUIER, Régine ANDRÉ-OBRECHT

Équipe SAMOVA, IRIT, UMR 5505 CNRS INPT UPS UT1
118, route de Narbonne, 31062 TOULOUSE Cedex 04, FRANCE
{pinquier,obrecht}@irit.fr

Résumé – Dans cet article, une nouvelle approche relative à l'indexation de la bonde sonore de documents audiovisuels est proposée, son but est de détecter et d'identifier des sons clés (jingles). La localisation de ces unités sonores permet, par exemple, de structurer le flux sonore en émissions (programmes). Chaque jingle, d'une longueur de une à quatre secondes ici, est représenté par une suite de vecteurs spectraux que nous nommerons "signature" par la suite. La détection de candidats potentiels est effectuée en comparant la signature de chacun des jingles au flux de données. Ce calcul de dissimilarité est réalisé avec la distance Euclidienne. Des règles heuristiques (basées sur des seuils) valident (confirment ou annulent) le choix des candidats potentiels préalablement sélectionnés. Afin de vérifier la faisabilité de notre système et de valider notre approche, des expériences sont réalisées sur des émissions télévisées et radiophoniques. Le volume de données, correspondant à trois chaînes de télévision et trois stations de radio, est de l'ordre d'une dizaine d'heures. Le système est efficace car les premiers résultats sont très encourageants. En effet, nous avons reconnu 130 jingles sur 132 avec un catalogue (tableau des jingles de référence) contenant 32 sons clés.

Abstract – This work addresses the soundtrack indexing of multimedia documents. Our purpose is to detect and locate one or many jingles to structure the audio dataflow in program broadcasts (reports). Each jingle, from one to four seconds length, is commonly represented by a sequence of spectral vectors, considered as its "signature". Potential candidates are extracted from the data flow by computing an Euclidean distance. They are validated with heuristic rules. The system evaluation is performed on TV and radio corpora (more than 10 hours, 3 TV channels and 3 radio channels). First results show that the system is efficient: among 132 jingles to recognize, we have detected 130 with our reference jingle table of 32 different key sounds.

1 Introduction

Le document audio ou la bande sonore d'un document audiovisuel est très complexe puisqu'il résulte d'un mixage entre plusieurs sources sonores. Si l'on se réfère à la norme MPEG7, indexer un document sonore signifie rechercher les composantes primaires (parole, musique), identifier des sons clés (applaudissements, effets spéciaux...), détecter et identifier les locuteurs, trouver des mots-clés ou rechercher des thèmes [1]. Néanmoins tous ces systèmes de détection présupposent l'extraction de composantes acoustiques élémentaires et homogènes. Dans la plupart des études, cette étape consiste à faire une discrimination parole/musique.

Plusieurs tendances sont observées. D'une part, dans la communauté des spécialistes en musique, l'accent porte sur des paramètres permettant de séparer au mieux la musique du reste (non-musique) [2]. Par exemple, le taux de passage par zéro (Zero Crossing rate) et le centroïde spectral sont utilisés pour séparer le bruit des parties voisées (donc harmoniques) [3], [4] tandis que la variation de la magnitude spectrale (le "Flux" spectral) permet de détecter les continuités harmoniques [5]. D'autre part, dans la communauté du traitement automatique de la parole, les paramètres cepstraux sont privilégiés pour extraire les zones de parole [6] et [7].

Dans une étude précédente [8], nous avons fusionné ces deux approches et obtenu de très bons résultats. En effet, il n'était plus question pour nous de chercher à discriminer la parole et la musique, mais à les caractériser au mieux de façon indépendante. Dans cet article, une alternative à ce premier partitionnement (parole/musique/autre) est proposée. Celle-ci consiste à détecter des sons-clés (appelés communément "jingles") représentant le début et/ou la fin d'une émission afin de segmenter ou structurer le flux audio-visuel [9]. Il ne s'agit pas de rechercher des thèmes [10], mais plutôt de proposer une macro-segmentation de l'audio en trouvant la structure temporelle des programmes télévisés ou radiophoniques.

La section "reconnaissance de sons" du document de spécifications de MPEG7 [11] propose une liste d'effets sonores classés en catégories afin de décrire les documents sonores. En effet, les sons clés de référence sont répertoriés dans un tableau dynamique : structurer un document sonore revient donc à détecter et localiser les occurrences de ces sons clés.

Notre étude se situe dans ce même cadre scientifique. Pour nous, un jingle est un extrait sonore qui dure généralement quelques secondes. Il a pour but de présenter le début ou la fin d'une émission (météo, journal, publicité...) ou d'attirer l'attention de l'auditeur. Celui-ci a la particularité de pouvoir aussi bien contenir de la musique que de la parole ou du bruit. Il est, de plus, généralement redondant. Nous appelons "jingle de référence", une occurrence de celui-ci. Ce descripteur audio "bas niveau" est basé sur une analyse spectrale. La distance Euclidienne est utilisée comme mesure de dissimilarité.

Cet article est divisé en deux parties. La première section présente le système global de classification qui permet de détecter et d'identifier les jingles présents dans le flux sonore à condition que ceux-ci fassent partis de notre catalogue de sons clés. La seconde section permet de valider notre approche par des expériences effectuées sur des documents audiovisuels et radiophoniques.

2 Le système de classification

Le système de classification (figure 1) est divisé en trois modules classiquement utilisés dans un problème de reconnaissance de formes :

- le prétraitement acoustique permet de caractériser au mieux le signal par une suite de vecteurs afin de comparer ceux-ci à la signature de chacun de nos jingles.
- la phase de détection propose des candidats potentiels issus de la comparaison par la distance Euclidienne.
- l'identification confirme ou annule le choix des candidats grâce à des heuristiques (seuils).

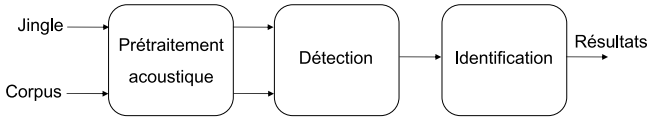

FIG. 1: *Le système global de détection et d'identification des jingles.*

2.1 Prétraitement acoustique

Ce prétraitement acoustique est basé sur une analyse spectrale (figure 2).

FIG. 2: *Extraction des paramètres par analyse spectrale.*

Le signal est découpé en trames de 32 ms avec recouvrement sur la moitié. Pour chaque trame d'analyse, une accentuation des aigus et un calcul du fenêtrage sont effectués (Hamming). Les coefficients spectraux sont alors créés à la suite du calcul des énergies dans les filtres par la FFT (Transformée de Fourier) et d'une pondération triangulaire (filtrage). Les filtres, couvrant la plage de fréquences [100 Hz - 8000 Hz], ont été testés lors d'une étude précédente sur la classification parole/musique [12].

Afin de ne pas tenir compte du facteur bruit/intensité qui peut varier au cours du temps ou des enregistrements, les spectres sont normalisés par leur énergie respective. Ainsi, 29 coefficients spectraux sont extraits.

2.2 Détection

Un jingle de référence, appartenant à notre catalogue de sons clés, est caractérisé par une suite de N vecteurs spectraux que nous appelons "signature" du jingle. Cette valeur N correspond au nombre de fenêtres d'analyse. La détection consiste à trouver cette séquence (suite de vecteurs) dans le flux de données à analyser. La distance Euclidienne est utilisée afin de comparer la signature (du jingle) et le signal (lui aussi représenté par une suite de vecteurs spectraux).

Cette comparaison s'effectue avec un pas de S vecteurs (figure 3).

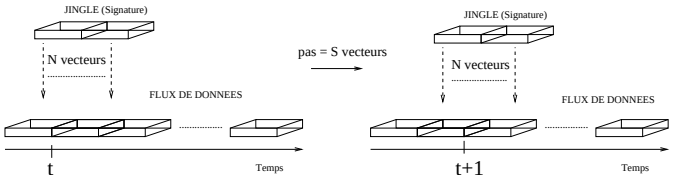

FIG. 3: *Comparaison entre le jingle et le corpus par distance Euclidienne.*

Les candidats potentiels sont sélectionnés comme étant des minima locaux. En effet, la valeur moyenne de la distance signature/flux est calculée. Si la distance courante est inférieure à la moitié de celle-ci, notée M dans la figure 5, cette distance est considérée comme une valeur minimale. Seuls les minima locaux, correspondants à ces valeurs minimales, sont détectés comme des jingles potentiels (figure 4).

2.3 Identification

La figure 4 est un exemple de résultats obtenus en calculant la distance Euclidienne entre la signature d'un jingle de référence et un fichier signal.

Nous pouvons observer cinq minima principaux qui ont été détectés dans l'étape précédente (cf. 2.2). Les deux premiers correspondent à un "bon" jingle : il s'agit d'un jingle présent dans le catalogue de sons-clés. Les trois autres sont bien des jingles mais n'appartiennent pas au catalogue. Ils ressemblent fortement aux deux premiers car les sons (notes) qui composent ces jingles sont les même mais passés dans un ordre différent.

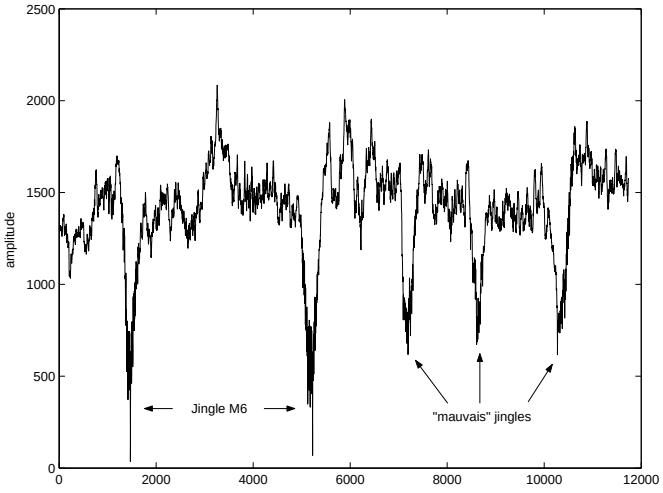

FIG. 4: *Distance Euclidienne lors de la détection du "jingle M6" sur 3 minutes du "corpus M6".*

Afin de sélectionner les "bons" jingles, nous proposons le processus suivant. Nous nous sommes aperçus que tous les minima correspondants à un jingle de référence, ont une particularité commune. En effet, ils sont représentés sans exception par un pic très fin. Ainsi, nous analysons la largeur des pics de chacun des minima locaux.

Pour cela, nous calculons (cf. figure 5) :

– h la valeur courante du minima local.

– L la largeur du pic à la hauteur H. H correspond à la hauteur à laquelle la largeur doit être estimée. Naturellement H et h sont liés (cf. 3.2).

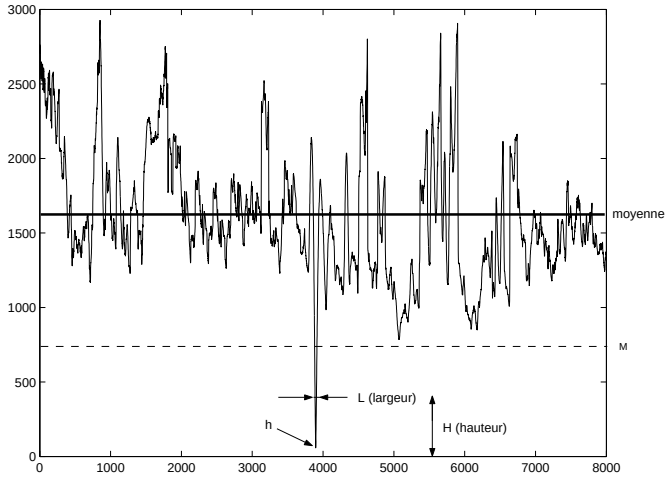

FIG. 5: *Identification des "bons" jingles par analyse de chacun des pics correspondant aux minima locaux détectés précédemment.*

Nous avons introduit un seuil λ. Si $L < \lambda$, le pic est fin et le minimum local est considéré comme un "bon" jingle. Sinon, le candidat est rejeté ("mauvais" jingle).

Remarque : le seuil λ doit être proportionnel à N.

3 Expériences

3.1 Corpus

Notre base de données est composée de six corpora différents (table 1). La durée totale est d'environ 10 heures. Cette base a été échantillonnée à 16 kHz.

TAB. 1: *Description de la base de données.*

Corpus	Durée	Jingles	Occurrences
France 3	15 min	1	4
M6	15 min	1	16
Canal+	30 min	1	6
France Info	60 min	1	12
RFI	360 min	3	60
Publicités	90 min	25	34
Total	**570 min**	**32**	**132**

- Le corpus "France Info" est composée d'émissions de radio et pour la plupart des informations composées d'actualités, de reportages, de sports et de la météo. Il y a aussi quelques chansons et des publicités.
- Le corpus "France 3" est un corpus télévisuel avec 2 chansons et diverses publicités.
- Les corpora "Canal+" et "M6" correspondent à des journaux d'informations de la télévision française.
- Le corpus "RFI" est multilingue (Français, Anglais et Espagnol) avec une majorité d'interviews, de reportages et d'informations de Radio France Internationale (RFI).
- Le dernier corpus est une compilation de plusieurs publicités télévisuelles et radiophoniques.

La durée d'un jingle varie de une à quatre secondes. Les jingles de références (ou sons clés) sont une sélection des jingles présents dans la base de données. Notre catalogue de sons clés est composé de 32 jingles différents. Plus de 200 jingles apparaissent dans notre base de données. Notre but est de détecter et d'identifier seulement les jingles identiques à ceux de notre catalogue de sons clés ou superposés à de la parole si le présentateur parle durant le jingle. Finalement, nous avons 132 jingles à reconnaître sur les 200.

3.2 Apprentissage

Afin d'implémenter notre méthode d'identification, nous devons tout d'abord fixer les paramètres S, H et λ. Pour cela, nous avons examiné le comportement de la distance Euclidienne entre les jingles de référence et le corpus France 3. Nous nous sommes aperçus que S pouvait être très grand sans pour autant que les résultats soient dégradés. Ce délai important (supérieur à une seconde pour un corpus échantillonné à 16 kHz) permet au système de fonctionner en temps réel.

Nous avons fixé H : $H = \alpha.h$, avec α correspondant à une constante ($\alpha = log2$).

Le seuil de rejet λ correspond au rapport entre la durée du jingle de référence N et le pas d'analyse S.

Experimentalement nous choisissons : $\lambda = 5 * N/S$.

3.3 Resultats

Nous avons testé **chaque** jingle de référence sur **tous** les corpora (table 2). Sur les 132 jingles que nous devions identifier, nous en avons détecté 130, soit 98,5% de taux de reconnaissance. Les deux seuls jingles omis (France Info et un jingle publicitaire) sont complètement recouverts de parole et leur pic est dans ce cas trop large (figure 4).

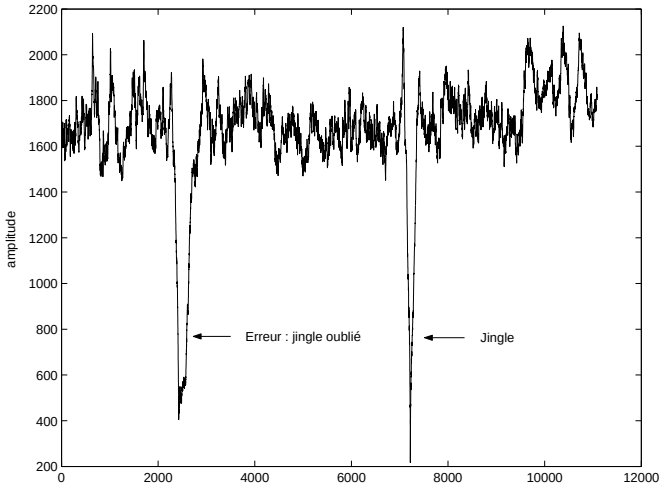

FIG. 6: *Exemple d'erreur (omission) du "jingle France Info" sur un extrait du corpus France Info (3 minutes).*

La détection est excellente : nous n'avons aucune fausse alarme et seulement deux omissions alors que d'autres jingles (n'appartenant pas au catalogue de sons clés) sont présents dans la base de données.

Bien que la base de données soit très variée, notamment par la différence des enregistrements entre les programmes de télévision et ceux de radio et la nature des émissions, le système possède un comportement très satisfaisant. En outre, ces expériences prouvent la robustesse de notre système.

TAB. 2: *Détection manuelle et automatique des jingles de référence sur chacun des corpus.*

Corpus	Détection auto	Détection manuelle
France 3	4	4
M6	16	16
Canal+	6	6
France Info	11	12
RFI	60	60
Publicités	33	34
Total	**130**	**132**

Durant la phase d'évaluation, nous avons étudié la précision de la détection. La localisation des jingles est très bonne : la différence entre les localisations manuelle et automatique sont très faibles (inférieur à 500 ms quelque soit le jingle). Dans une optique d'indexation, où généralement les décisions sont prises pour chaque seconde d'analyse, cette précision est suffisante.

Nous pouvons aussi noter que notre système fonctionne en temps réel. En effet, pour traiter un fichier sonore d'une heure avec notre catalogue de 32 jingles de référence, moins d'une heure est nécessaire en utilisant un processeur AMD cadencé à 1,4 GHz.

4 Discussion

Dans cet article, une nouvelle approche relative à la classification de la bande sonore est proposée, son but est d'indexer les documents sonores. Nous avons présenté pour cela un système de détection et d'identification de jingles (ou sons clés) basé sur un calcul de distance euclidienne dans le domaine spectral. Cette méthode est assez simple mais, néanmoins les résultats sont excellents. En effet, nous n'observons aucune fausse alarme et seulement deux omissions dans des conditions extrêmes : de la parole est superposée au jingle pendant l'**intégralité** de celui-ci. La localisation est très satisfaisante : nous pouvons déterminer le début d'un jingle avec une marge inférieure à la demi-seconde, ce qui est largement suffisant pour une tâche d'indexation.

Notre système peut être considéré comme efficace par sa simplicité (seulement basé sur une analyse spectrale), sa rapidité (temps réel), sa robustesse (indépendant du corpus) et la qualité de ses résultats. Il peut être utilisé pour une description des documents sonore de plus haut niveau, de manière à par exemple structurer ou classer en émissions (programmes).

Ce travail devra être prolongé en y ajoutant une macro segmentation visuelle [13] afin de définir une signature audio/vidéo des émissions et trouver une mesure de dissimilarité audiovisuelle.

Références

[1] M. Franz, J. Scott McCarley, T. Ward, and W. Zhu, "Topics styles in IR and TDT: Effect on System Behavior," in *European Conference on Speech Communication and Technology*, Aalborg, Denmark, Sept. 2001, pp. 287–290.

[2] S. Rossignol, X. Rodet, J. Soumagne, J. L. Collette, and P. Depalle, "Automatic Characterization of Musical Signals: Feature Extraction and Temporal Segmentation," *Journal of New Music Research*, vol. 28, no. 4, pp. 281–295, Dec. 1999.

[3] J. Saunders, "Real-time Discrimination of Broadcast Speech/Music," in *IEEE International Conference on Audio, Speech and Signal Processing*, Atlanta, USA, May 1996, pp. 993–996.

[4] T. Zhang, C. Kuo, and C. J., "Hierarchical System for Content-Based Audio Classification and Retrieval," in *Conference on Multimedia Storage and Archiving Systems III*, Nov. 1998, vol. 3527 of *SPIE*, pp. 398–409.

[5] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," in *IEEE International Conference on Audio, Speech and Signal Processing*, Munich, Germany, Apr. 1997, pp. 1331–1334.

[6] J. L. Gauvain, L. Lamel, and G. Adda, "Systèmes de processus légers : concepts et exemples," in *International Workshop on Content-Based Multimedia Indexing*, Toulouse, France, Oct. 1999, pp. 67–73, GDR-PRC ISIS.

[7] J. Foote, "Automatic Audio Segmentation using a Measure of Audio Novelty," in *IEEE International Conference on Multimedia and Expo*, New-York, USA, 2000, pp. 452–455.

[8] J. Pinquier, Jean-Luc Rouas, and R. André-Obrecht, "A Fusion Study in Speech / Music Classification," in *IEEE International Conference on Audio, Speech and Signal Processing*, Hong-Kong, China, Apr. 2003.

[9] J. Carrive, F. Pachet, and R. Ronfard, "CLAViS - A Temporal Reasoning System for Classification of Audiovisual Sequences," in *Content-Based Multimedia Information Access Conference (RIAO)*, College de France, Paris, France, Apr. 2000.

[10] R. Amaral, T. Langlois, H. Meinedo, J. Neto, N. Souto, and I. Trancoso, "The Development of a Portuguese Version of a Media Watch System," in *European Conference on Speech Communication and Technology*, Aalborg, Denmark, Sept. 2001, vol. 4, pp. 2689–2692.

[11] ANSI, "ISO/IEC 15938-4 Information Technology - Multimedia Content Description Interface - Audio," Tech. Rep., MPEG, 2001.

[12] J. Pinquier, C. Sénac, and R. André-Obrecht, "Indexation de la bande sonore : recherche des composantes parole et musique," in *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, Angers, France, Jan. 2002, pp. 163–170.

[13] P. Aigrain, P. Joly, and V. Longueville, "Medium Knowledge-Based Macro-Segmentation of Video into Sequences," *In Intelligent Multimedia Information Retrieval*, pp. 159–173, 1997.