Structuration audiovisuelle par composantes sonores primaires

Julien PINQUIER et Régine ANDRÉ-OBRECHT

Équipe SAMOVA - IRIT - UMR 5505 CNRS INP UPS
118, route de Narbonne - 31062 Toulouse cedex 04, FRANCE
{pinquier, obrecht}@irit.fr

Résumé : Le développement croissant des données numériques et l'explosion des accès multimédia à l'information, sont confrontés au manque d'outils automatiques efficaces. Dans ce cadre, plusieurs approches relatives à l'indexation et la structuration de la bande sonore de documents audiovisuels sont proposées. Leurs buts sont de détecter les composantes primaires telles que la parole, la musique et les sons clés (jingles, sons caractéristiques, mots clés...). Pour la classification parole/musique, quatre paramètres originaux sont extraits : la modulation de l'entropie, la modulation de l'énergie à quatre hertz, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde. Des expériences sur un corpus radiophonique montrent la robustesse de ces paramètres : notre système possède un taux de classification correcte supérieur à 90 %. Un autre partitionnement consiste à détecter des sons clés. La sélection de candidats potentiels est effectuée en comparant la « signature » de chacun des jingles au flux de données. Ce système est simple par sa mise en œuvre mais rapide et très efficace : sur un corpus audiovisuel d'une dizaine d'heures (environ 200 jingles) aucune fausse alarme n'est présente. Il y a seulement deux omissions dans des conditions extrêmes. Les applaudissements sont modélisés à l'aide de MMG dans le domaine spectral. Un corpus télévisuel permet de valider cette première étude par des résultats encourageants. Grâce à l'extraction de ces composantes primaires, les émissions audiovisuelles peuvent être annotées de manière automatique. Une réflexion est conduite quant à l'utilisation de ces composantes afin de trouver une structure temporelle à nos documents : il s'agit de détecter un motif récurrent dans une collection d'émissions, dites de plateau.

Mots-clés : indexation sonore, structuration audiovisuelle, classification, énergie, entropie, segmentation, parole, musique, jingles, applaudissements.

1 Introduction

Par analogie avec les documents textuels qui sont faciles à manipuler (stockage, manipulation et recherche d'information étant devenus des opérations abordables par le grand public), le traitement des documents multimédia n'en est qu'à son balbutiement. Par exemple, trouver la vidéo contenant les premiers pas d'Armstrong sur la Lune (sans information *a priori*) est pour l'instant assez critique si l'on ne traite que les documents multimédia. Il serait souhaitable, comme en indexation textuelle, que l'on puisse utiliser des moteurs de recherche via des mots clés. Cela nécessite d'extraire du sens de la vidéo et/ou de l'audio et de les utiliser conjointement.

Un document sonore, c'est-à-dire la bande sonore d'un document multimédia ou enregistrement d'émission radiophonique, est un document particulièrement difficile à indexer, car l'extraction de l'information élémentaire se heurte à l'extrême diversité des sources acoustiques. Les segments acoustiques sont de nature très diverses de par leur production et leur enregistrement : l'environnement peut être propre ou plus ou moins bruité, la qualité de l'enregistrement peut être plus ou moins soignée et liée à des éléments extérieurs (canal téléphonique), la musique peut être traditionnelle ou synthétique, la présence de parole peut être observée en monologue ou en dialogue...

Il peut être intéressant de rechercher des « bruits » ou des sons sémantiquement significatifs tels que les applaudissements, les rires ou les effets spéciaux (pistolets, explosions...), de repérer les passages musicaux pour les segmenter et les identifier, de détecter les locuteurs équivalents à des tours de parole dans un dialogue. Si l'on se réfère à la norme MPEG7, indexer un document sonore signifie rechercher aussi bien des composantes de bas niveau dites primaires comme la parole, la musique, les sons clés (jingles, mots-clés...) que des descripteurs de plus haut niveau tels les locuteurs ou les thèmes.

Dans cet article, nous présentons un système de détection de composantes primaires telle la parole et la musique. Ce système est fondé sur l'extraction de trois paramètres originaux : la modulation de l'entropie, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde. Les informations issues de ces paramètres sont ensuite fusionnées avec celle issue de la modulation de l'énergie à quatre hertz. Au delà du partitionnement primaire, il est intéressant de détecter des sons clés ou des jingles représentant le début et/ou la fin d'un segment sonore afin de structurer le flux audio-visuel (Carrive *et al.*, 2000). Il ne s'agit pas de rechercher des thèmes (Amaral *et al.*, 2001), mais plutôt de proposer une macrosegmentation de l'audio en trouvant sa structure temporelle. Nous décrivons nos systèmes de détection de jingles et d'applaudissements. Ensuite, pour chacun des systèmes, nous proposons des expériences sur des corpora radiophoniques et télévisuels. Enfin, nous présentons deux exemples de fusion possible de nos outils en vue d'une structuration de plus haut niveau : il s'agit d'une recherche de motif dans une collection d'émissions et d'une segmentation d'un journal télévisé.

2 Système de détection parole/musique

Le système se décompose en deux systèmes de classification correspondant aux deux détections disjointes de la parole et de la musique. Il est fondé sur l'extraction de quatre paramètres (cf. figure 1) : la modulation de l'énergie à 4 Hertz, la modulation de l'entropie, le nombre de segments par seconde et la durée de ces segments (Pinquier *et al.*, 2003). La décision est prise en comparant les scores (vraisemblances) issus de la modélisation de chacun des paramètres considérés.

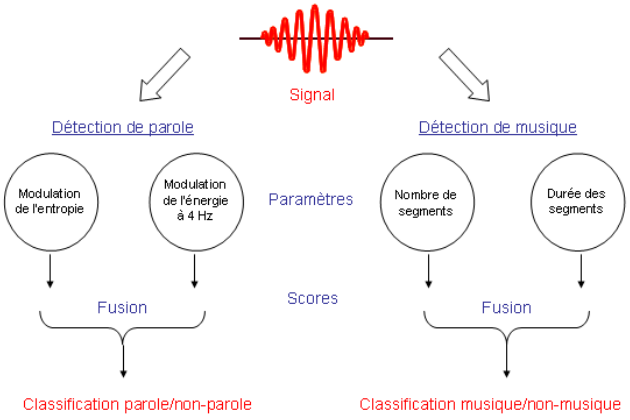

FIG. 1 – Le système de classification parole/musique.

2.1 Détection de parole

La détection des zones de parole est effectuée à partir de la fusion des deux paramètres de modulation.

2.1.1 Modulation de l'énergie à 4 Hertz

Le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4 Hertz (Houtgast & Steeneken, 1985). En effet, ces modulations correspondent au rythme syllabique. La parole possède une modulation de l'énergie à 4 Hertz plus forte que la musique.

2.1.2 Modulation de l'entropie

Des observations menées sur le signal ainsi que sur le spectrogramme font apparaître une structure plus « ordonnée » du signal de musique que de parole. Pour mesurer ce « désordre », nous avons calculé un paramètre fondé sur l'entropie du signal (Moddemeijer, 1989). La modulation de l'entropie est alors plus élevée pour la parole que pour la musique (cf. figure 2).

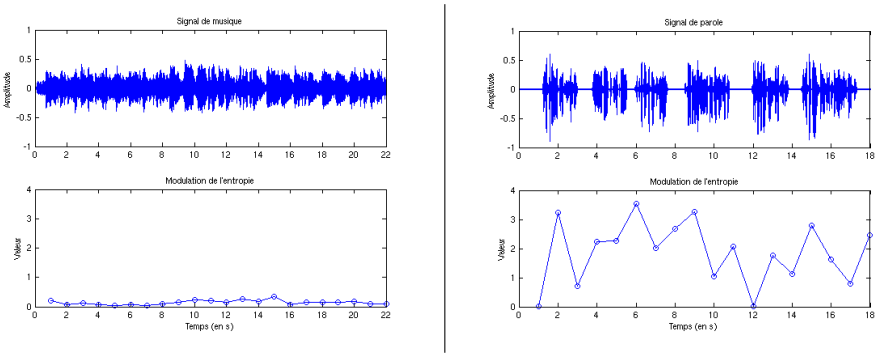

FIG. 2 – Modulation de l'entropie pour la musique (extrait de Mozart de 22 s) et la parole (6 phrases de parole lue de 18 s).

2.2 Détection de musique

La détection des zones de musique est réalisée grâce à deux paramètres issus d'une segmentation automatique du signal. La longueur des segments quasi stationnaires est différente pour la parole et la musique. En utilisant une segmentation du signal en zones quasi stationnaires, nous cherchons à mettre en évidence cette information. La segmentation est issue de l'algorithme de « Divergence Forward-Backward » (DFB) (André-Obrecht, 1988) qui est fondé sur une étude statistique du signal dans le domaine temporel.

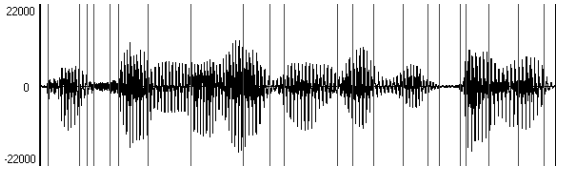

FIG. 3 – Résultat de la segmentation sur environ 1 seconde de parole. La phrase prononcée est : « Confirmez le rendez-vous par écrit ».

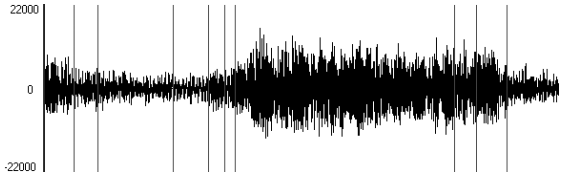

FIG. 4 – Résultat de la segmentation sur environ 1 seconde de musique d'un extrait de Mozart.

2.2.1 Nombre de segments

Le nombre de segments présents durant chaque seconde de signal est calculé. Les signaux de parole présentent une alternance de périodes de transition (voisées/non-voisées) et de périodes de relative stabilité (les voyelles en général) (Calliope, 1989). Au niveau de la segmentation, cela se traduit par de nombreux changements. La musique, étant plus tonale (ou harmonique), ne présente pas de telles variations. Le nombre de segments par unité de temps (ici la seconde) est donc plus important pour la parole que pour la musique.

2.2.2 Durée des segments

La durée des segments, obtenue après segmentation automatique (DFB), est fortement corrélée au nombre de segments par seconde. Afin de limiter la corrélation de ces deux paramètres de segmentation, la durée moyenne des segments sur une seconde est calculée sur les 7 segments les plus longs de la seconde. Le nombre de segments caractéristiques est fixé expérimentalement. Les segments sont généralement plus longs pour la musique que pour la parole.

3 Système de détection de jingles

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. Il a la particularité de pouvoir aussi bien contenir de la musique que de la parole. Il est, de plus, généralement redondant dans une collection de documents audiovisuels. Le système de détection d'un jingle est divisé en trois modules classiquement utilisés dans un problème de reconnaissance de formes (cf. figure 5).

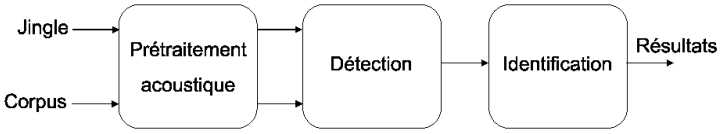

FIG. 5 – Schéma général de détection d'un jingle.

3.1 Prétraitement acoustique

Le pré-traitement acoustique est fondé sur une analyse spectrale. Le signal est découpé en trames de 32 ms avec recouvrement sur la moitié de la trame. 28 coefficients spectraux sont extraits (Pinquier *et al.*, 2002).

3.2 Détection et identification

Un jingle de référence est caractérisé par une suite de N vecteurs spectraux que nous appelons « signature ». Cette valeur correspond au nombre de fenêtres d'analyse obtenues sur la durée totale du jingle considéré. 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. Cette comparaison s'effectue donc sur une fenêtre glissante de vecteurs que l'on déplace par un pas de S vecteurs (cf. figure 6).

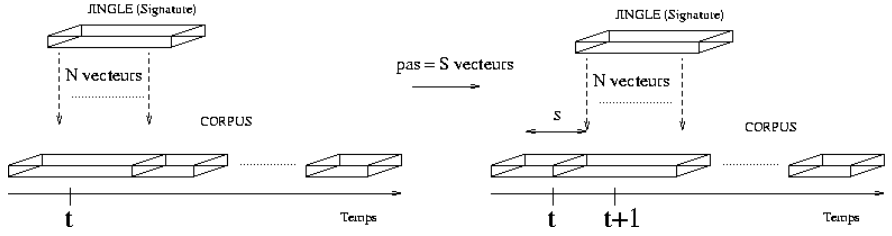

FIG. 6 – Comparaison entre le jingle et le corpus par distance Euclidienne.

Les candidats potentiels correspondent à certains minima locaux de la distance signature/flux calculée. L'analyse proposée consiste à étudier la largeur des pics de chacun des minima locaux à l'aide des variables de valeur du minimum local h, de hauteur relative H et de largeur de pic L (cf. figure 7).

Ce système est plus détaillé dans (Pinquier & André-Obrecht, 2004).

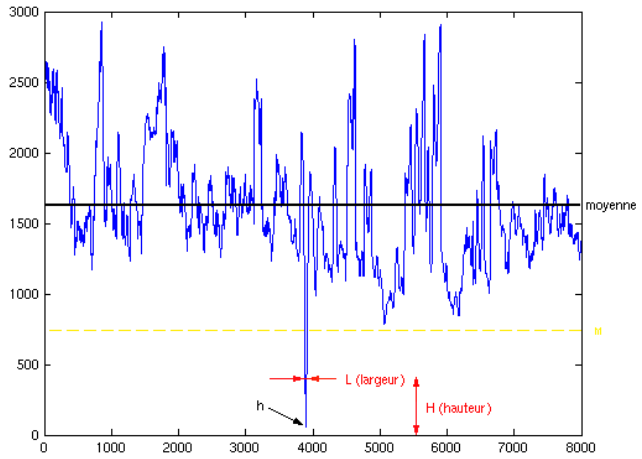

FIG. 7 – Sélection des jingles.

4 Détection des applaudissements

Le système est composé de deux modules principaux : le prétraitement et la décision (ou reconnaissance). Le prétraitement correspond à une analyse spectrale (voir section 3.1). Le système de détection mis en place s'inspire très largement des systèmes classiques parole/musique (Gauvain *et al.*, 1999). Il consiste à identifier sur chaque trame de signal, la présence ou l'absence du phénomène considéré en question (parole, musique, applaudissements...). Il s'agit d'un problème de classification en classe (applaudissements) et non-classe (non-applaudissements). La modélisation est effectuée à l'aide de Modèles de Mélanges de lois Gaussiennes et un apprentissage est alors nécessaire.

4.1 Reconnaissance

La décision se fait suivant la règle du maximum de vraisemblance entre les modèles applaudissements et non-applaudissements. Une fonction de lissage permet de ne garder que les segments significatifs (représentatifs d'une zone d'applaudissements), après regroupement des trames correspondant à la même décision. Le lissage est d'une seconde.

4.2 Apprentissage

L'apprentissage de ces paramètres est classiquement réalisé par les algorithmes VQ (Lloyd, 1957) pour l'initialisation des modèles et EM (Dempster *et al.*, 1977) pour la réestimation et l'optimisation des paramètres du mélange.

Après expérimentations, le nombre de lois gaussiennes de chacun des modèles a été fixé à 64 (cf. figure 8).

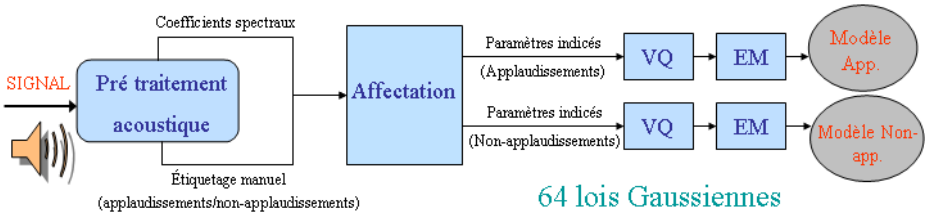

FIG. 8 – Apprentissage des modèles de mélanges de lois gaussiennes représentant les applaudissements et les non-applaudissements.

5 Expériences

5.1 Corpus

Notre base de données est assez hétérogène. L'apprentissage des seuils du système de classification parole/musique a été effectué sur 3 heures de parole lue : corpus MULTEXT (Campione & Véronis, 1998) et sur 3 heures d'extraits musicaux. Les tests sont effectués sur des données radiophoniques et télévisuelles pour une durée totale d'environ 12 heures.

Beaucoup de jingles apparaissent dans notre base de données. Notre but est de détecter et d'identifier seulement les jingles similaires à ceux de notre catalogue de sons clés. Nous avons 132 jingles à retrouver et reconnaître, sachant que notre catalogue est composé de 32 jingles de référence.

Pour la détection des applaudissements, nous utilisons le corpus « Le Grand Échiquier ». Le contenu de ce corpus est assez divers : de la musique (classique, jazz, variété française...), des interviews et des sketches. Chaque émission a une durée d'environ 190 minutes. La première que nous appelons « GE1 » nous sert d'apprentissage et la seconde « GE2 » nous sert de test. Nous utilisons donc au total seulement 6 heures de notre corpus. L'utilisation d'autres émissions est possible, des tests ont d'ailleurs été effectués sur deux autres émissions. La tâche d'annotation manuelle nécessaire pour évaluer les résultats étant très pénible, nous avons fait cet étiquetage que pour deux émissions seulement.

5.2 Evaluation

5.2.1 Classification parole/musique

Chaque paramètre est pertinent dans le sens où il permet, en lui-même, de faire une discrimination parole/non-parole ou musique/non-musique correcte. En considérant chaque paramètre individuellement, le taux de classification correcte varie d'environ 78 % pour la durée des segments à plus de 87 % pour la modulation de l'entropie (Table 1).

TAB. 1 – *Résultats de la classification parole/musique.*

Paramètres	Taux de classification correcte
(1) Modulation de l'énergie à 4 Hz	87.3 %
(2) Modulation de l'entropie	87.5 %
(3) Nombre de segments	86.4 %
(4) Durée des segments	78.1 %
(1) + (2) Détection de parole	**90.5 %**
(3) + (4) Détection de musique	**89 %**

La fusion entre ces paramètres par maximisation des scores de vraisemblances permet d'améliorer les résultats et d'obtenir environ 90 % de reconnaissance correcte pour chacune des classifications (parole/non-parole et musique/non-musique).

5.2.2 Détection de jingles

Sur les 132 jingles que nous devions localiser et identifier, nous en avons détecté 130, soit 98,5 % de taux de reconnaissance. Les deux seuls jingles omis (un jingle « France Info » et un jingle publicitaire) sont complètement recouverts de parole (le présentateur parle durant le jingle !) et leur pic est dans ce cas beaucoup trop large. La détection est excellente car nous n'avons aucune fausse alarme (insertion) et seulement deux omissions alors que d'autres jingles n'appartenant pas au catalogue de sons clés sont présents dans la base de données.

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 est très faible, inférieure à 500 ms quel que soit le jingle ; elle correspond au pas S utilisé.

5.2.3 Détection des applaudissements

Les résultats de la détection des applaudissements sont intéressants : nous obtenons 98,58 % de taux de classification correcte. Ce taux est à relativiser : il nous faut observer les applaudissements retrouvés. Sur les 906 secondes d'applaudissements que nous avons repérées manuellement, nous avons relevé 144 segments. Seulement 72 de ces segments sont significatifs et ceux-ci sont tous

bien détectés par notre système. Lorsque nous employons le terme « significatif », il désigne des *segments assez longs*, de durée supérieure à 1 seconde, et *pur* : ce ne sont pas des segments de faibles amplitudes ou superposés à de la parole. Le système de détection des applaudissements est excellent car il n'y a pratiquement pas d'insertions et tous les segments significatifs sont retrouvés.

6 Structuration

Nous nous plaçons ici dans le cadre d'une première analyse de scène par les composantes sonores primaires : la recherche de parole, de musique, de jingles et d'applaudissements. Ces informations de « bas niveau », extraites directement du flux sonore, ne sont pas directement exploitables pour la structuration de documents audiovisuels. Pour accéder à une information de plus haut niveau, il faut d'une part les regrouper, et d'autre part voir leurs impacts sur les autres informations sonores. Nous proposons un exemple de fusion possible de nos outils de segmentation sonore en vue d'une structuration de plus haut niveau : une recherche de motif sur une collection d'émissions.

6.1 Détection de motif dans une collection d'émissions

6.1.1 Présentation

Lorsqu'une seule émission est analysée, un traitement très spécifique peut être effectué. Suivant la durée de l'émission, un traitement manuel peut même être réalisé et s'avérer plus rapide. Par contre, quand nous sommes en présence de plusieurs émissions, comme la collection du « Grand Échiquier » présentée dans la section 5.1 dont nous possédons 54 émissions, il est nécessaire de changer de stratégie. Les traitements ne peuvent être qu'automatiques vue la durée du corpus (plus de 160 heures). Le niveau structurel doit rester assez grossier afin de correspondre à toute la collection d'émissions considérée.

L'étude de cette collection a permis de définir un motif, c'est-à-dire un enchaînement récurrent de caractéristiques communes à chacune des émissions. Ce motif structure les émissions en parties homogènes ; en général, il s'agit du passage d'un invité à un autre. Le motif est le suivant :
présentateur / *[applaudissements]* / **spectacle** / *[applaudissements / spectacle]* / **applaudissements** / **présentateur**.

Ceci signifie que dans cette collection, un spectacle (chanson, morceau de musique, sketch, extrait de film...) est introduit par le présentateur et est suivi par des applaudissements. À la fin de ceux-ci, le présentateur reprend la parole. Des applaudissements peuvent éventuellement précéder la composition artistique ou la « découper ».

Nous avons choisi de rechercher cette structuration sur le même fichier de test que précédemment pour des raisons évidentes : nous possédons déjà la « vérité terrain » correspondant aux détections d'applaudissements et du présentateur pour cette émission. Lors de cette émission, nous avons répertorié une succession de dix de ces motifs. Afin de retrouver le motif en question, de manière automatique, nous allons appliquer une stratégie dite « aveugle ».

Trois classifications sont effectuées indépendamment les unes des autres (cf. figure 9) :
- une détection de musique permettant de repérer les chansons,
- une détection de parole, pré-traitement pour la recherche du présentateur,
- une détection des applaudissements.

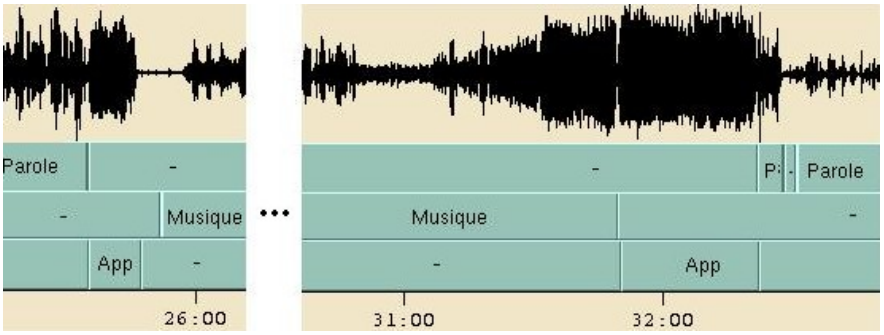

FIG. 9 – Exemple de recherche de motif sur 7 minutes de l'émission « GE2 » de la collection du « Grand Échiquier » à travers les détections automatiques de parole (ligne 1), de musique (ligne 2) et d'applaudissements (ligne 3).

Notons, que les modèles des applaudissements et des non-applaudissements sont les mêmes que ceux développés précédemment. Ils ont été appris sur une autre émission : « GE1 ». Rappelons aussi, que pour les détections de parole et de musique, notre système ne nécessite pas d'apprentissage. Il n'y a donc eu aucun apprentissage sur GE2.

La figure 10 est un exemple de résultat obtenu par une mise en commun de tous les résultats de détection.

Les résultats sont excellents sur l'émission « GE2 ». Les 10 motifs cherchés sont retrouvés et ceci bien que les détections de parole et de musique soient imparfaites : quelques légers décalages et des insertions de parole sur la musique sont observés.

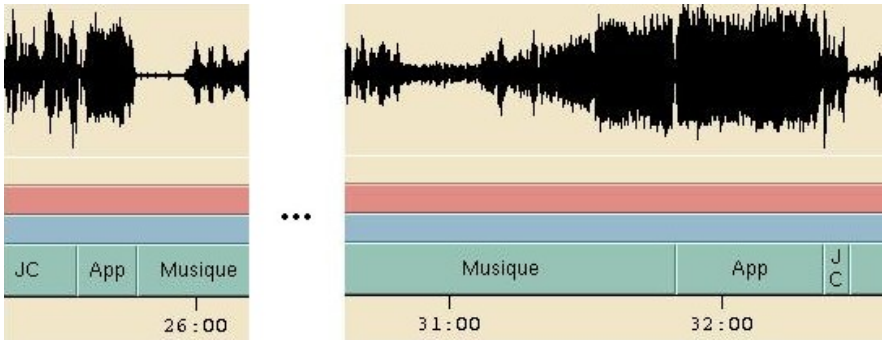

FIG. 10 – Exemple de résultat obtenu par fusion des différentes détections sur un extrait de 7 minutes de l'émission « GE2 » du « Grand Échiquier ». « JC » représente Jacques Chancel.

6.1.2 Remarques

Afin de valider totalement notre approche, l'ensemble de cette structuration va être effectuée sur l'ensemble des 54 émissions de cette collection sans aucun autre apprentissage ou intervention manuelle. Comme nous l'avons déjà signalé, le motif recherché n'est pas spécifique au « Grand Échiquier » mais commun à nombre d'émissions de plateaux.

La recherche d'un tel motif sur une autre émission ne nécessite alors qu'un étiquetage de 2 ou 3 minutes du nouveau présentateur de l'émission afin de créer son propre modèle acoustique. Ceci est effectué par adaptation du modèle dit du « monde » Les autres détections ne nécessitent pas d'apprentissage.

Le choix de ce motif s'est effectué après une intervention humaine, à savoir l'écoute de deux émissions du « Grand Échiquier ». Il s'est avéré que ce motif a pu être retrouvé tout au long des émissions et a été ainsi validé.

Il serait intéressant de pouvoir trouver de manière automatique le motif récurrent d'une émission. Le principe pourrait être le suivant :

- annotation automatique à partir des outils d'analyse audio et vidéo,
- recherche automatique des suites récurrentes dans la succession des annotations,
- inférence d'un motif,
- structuration du document à partir du motif trouvé.

La deuxième étape s'apparente à une détection d'invariants audiovisuels et fait l'objet de recherche récentes Haidar *et al.* (2004).

7 Conclusion

Dans le contexte de l'indexation sonore, nous avons étudié différentes composantes primaires, permettant une structuration audiovisuelle. Pour chacune de ces unités bas niveau, un détecteur automatique est développé afin de les extraire du continuum sonore.

Pour les spécialistes de l'audio, les composantes primaires correspondent souvent à la parole et à la musique. Le système original que nous avons développé est fondé sur une fusion de quatre paramètres : la modulation de l'énergie à 4 Hertz, la modulation de l'entropie, le nombre de segments issus d'une segmentation automatique et la durée de ces mêmes segments. Les résultats obtenus sont très bons, plus de 90 % de classification correcte, mais surtout le système est très robuste : il ne nécessite aucun nouvel apprentissage et/ou adaptation de ses seuils contrairement aux approches classiques fondées sur une analyse spectrale et des modèles de mélanges de lois gaussiennes.

D'autres composantes primaires correspondent à des sons clés : nous avons étudié les jingles et les applaudissements. Notre détecteur de jingles est excellent. Bien que la méthode soit assez simple, mesure de distance dans le domaine spectral, sur 10 heures de tests et 132 jingles à retrouver, nous n'avons observé que 2 omissions et aucune fausse alarme. Les erreurs apparaissent dans des conditions très particulières, par exemple là où la parole recouvre entièrement le jingle. Cette étude est d'autant plus intéressante qu'elle ne nécessite aucun apprentissage, seulement une occurrence du jingle à reconnaître. Une étude sur la détection des applaudissements a permis de montrer la faisabilité d'une méthode fondée sur une analyse spectrale et des modèles de mélanges de lois gaussiennes. Les résultats sont excellents car les sons caractéristiques, utiles en vue d'une structuration, sont tous détectés.

À la suite des détections de ces différentes composantes primaires, une étude en structuration a été effectuée : il s'agit d'une détection de motif sur une collection d'émissions. Elle permet de mettre en parallèle nos différentes détections afin d'extraire un enchaînement récurrent dans des émissions issues de la collection le « Grand Échiquier ». Il s'agit de la séquence : présentateur / [applaudissements] / spectacle / [applaudissements / spectacle] / applaudissements / présentateur. Durant notre émission de test, les 10 occurrences de ce motif sont toutes détectées.

Ces premiers travaux de structuration sont encourageants, mais il est fort dommage de se limiter à l'analyse d'un seul media (le son dans notre cas), alors que nous exploitons des bases de données audiovisuelles. C'est pourquoi nous devrons réfléchir sur l'apport de l'analyse vidéo. La détection de logos et la reconnaissance de l'intervenant sont des études complémentaires à la détection de jingles et à la reconnaissance de locuteur.

Ces travaux sont actuellement en cours au sein de notre équipe sur l'analyse de la vidéo.

Cette immersion dans l'analyse de la vidéo doit nous permettre de mieux cerner les complémentarités entre l'audio et la vidéo et d'appréhender à moyen terme le vrai problème du traitement audiovisuel. Il est impératif de savoir répondre plus précisément à des questions classiques du type :

- qu'est ce qu'une information audiovisuelle ? Qu'est ce qu'une indexation audiovisuelle ? La présence d'un personnage est une illustration simple de ce type d'information, voire d'index. Peut-on généraliser cette démarche ?
- qu'est ce qu'une analyse audiovisuelle ? Sachant que, comme nous avons essayé de le montrer, une analyse audiovisuelle ne signifie pas une simple fusion d'informations issues d'une analyse audio et d'une analyse vidéo.

Références

Amaral R., Langlois T., Meinedo H., Neto J., Souto N. & Trancoso I. (2001). The Development of a Portuguese Version of a Media Watch System. In *European Conference on Speech Communication and Technology*, volume 4, p. 2689–2692, Aalborg, Denmark.

André-Obrecht R. (1988). A New Statistical Approach for Automatic Speech Segmentation. *IEEE Transactions on Audio, Speech, and Signal Processing*, **36**(1), 29–40.

Calliope (1989). *La parole et son traitement automatique*. Paris, France : Masson.

Campione E. & Véronis J. (1998). A Multilingual Prosodic Database. In *International Conference on Spoken Language Processing*, p. 3163–3166, Sydney, Australia.

Carrive J., Pachet F. & Ronfard R. (2000). CLAViS - A Temporal Reasoning System for Classification of Audiovisual Sequences. In *Content-Based Multimedia Information Access Conference (RIAO)*, College de France, Paris, France.

Dempster A. P., Laird N. M. & Rubin D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39 (Series B)**, 1–38.

Gauvain J. L., Lamel L. & Adda G. (1999). Systèmes de processus légers : concepts et exemples. In *International Workshop on Content-Based Multimedia Indexing*, p. 67–73, Toulouse, France : GDR-PRC ISIS.

Haidar S., Joly P. & Chebaro B. (2004). Detection Algorithm of Audiovisual Production Invariant. In *Workshop on Adaptative Multimedia Retrieval (AMR)*, p. 156–169, Valencia, Spain.

Houtgast T. & Steeneken J. M. (1985). A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria. *Journal of the Acoustical Society of America*, **77**(3), 1069–1077.

Lloyd S. P. (1957). Least Squares Quantization in PCM. Unpublished Bell Labs Technical Note (1957).

Moddemeijer R. (1989). On Estimation of Entropy and Mutual Information of Continuous Distributions. *Signal Processing*, **16**(3), 233–246.

Pinquier J. & André-Obrecht R. (2004). Jingle detection and identification in audio documents. In *International Conference on Audio, Speech and Signal Processing*, Montréal, Canada.

Pinquier J., Rouas J.-L. & André-Obrecht R. (2003). Fusion de paramètres pour une classification automatique parole/musique robuste. *Technique et Science Informatiques (TSI)*, **22**(7-8), 831–852.

Pinquier J., Sénac C. & André-Obrecht R. (2002). Indexation de la bande sonore : recherche des composantes parole et musique. In *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, p. 163–170, Angers, France.