CYNDI : un moteur d'indexation de la bande sonore par une segmentation sémantique et extraction de mots clés

Hadi Harb[1] Liming Chen[1]

[1] Laboratoire LIRIS, Département Mathématiques Informatique

Ecole Centrale de Lyon
36, av. Guy de Collongue, 69134, Ecully, France

{hadi.harb, liming.chen}@ec-lyon.fr

Résumé

La multiplication de Documents Audio-Visuels nécessite des outils de recherche et de navigation pour faire face à la profusion de ceux-ci. Dans cet article nous présentons CYNDI, notre moteur d'indexation de la bande sonore de programmes audiovisuels que nous avons réalisé dans le cadre du projet RNRT CYRANO. Notre moteur d'indexation CYNDI s'appuie sur une segmentation automatique de la bande sonore en musique ou parole, puis pour les segments de parole une technique de segmentation en « phrases » qui permet de faciliter la navigation dans un document audiovisuel tout en diminuant le taux d'erreur d'un moteur de transcription automatique. Notre technique de segmentation en phrases à partir d'un segment de parole est basée sur des statistiques de la taille normale d'une phrase. La détection de bordures de phrases est basée sur un seuillage automatique des valeurs de la distance de KullBack-Leibler. Les segments de musique sont indexés par CYNDI d'une manière à permettre une recherche par similarité.

Mots clefs

Indexation audio, MPEG7, indexation par le contenu, Accès intelligent aux vidéos.

1 Introduction

Le projet RNRT CYRANO est un projet associant France Télécom R&D et l'INRIA pour un accès coopératif et personnalisé de documents multimédias sur les grands réseaux. L'objectif du projet est de développer des outils pour faire face à la profusion de documents multimédias.

Dans le cadre du projet Cyrano, nous avons travaillé sur l'indexation par le contenu de programmes audiovisuels que sont par exemple les journaux télévisées, les meetings et les documentaires. Il s'agit de créer ou générer automatiquement des index sémantiques facilitant la navigation et la recherche au sein d'un programme audiovisuel. Dans ce papier, nous nous intéressons à l'indexation sémantique de la bande sonore qui est associée à tout programme audiovisuel. L'objectif ici est la création automatique d'un index capturant le découpage d'une bande sonore en des segments de parole ou de musique. Puis pour les segments de parole la structure de « phrase » et enfin des mots clés résumant le contenu de chaque «phrase ». Les segments de musique sont ensuite indexés par le contenu afin de permettre une recherche par similarité de ces segments.

Dans nos travaux, la génération de mots clés fait appel à un moteur de Reconnaissance Automatique de la Parole (RAP), en occurrence ViaVoice d'IBM. Nos expérimentations montrent qu'une application brute d'un RAP sur un flux sonore en continu conduit à une diminution sensible du taux de reconnaissance du fait de la discontinuité linguistique entre les phrases ou paragraphes. Une étape préalable de la segmentation de la bande sonore en des segments homogènes et de taille acceptable est donc indispensable, d'autant plus qu'une telle segmentation permet une navigation intelligente pour faire des sauts de paragraphe.
CYNDI utilise une nouvelle technique pour l'indexation par le contenu de la musique basée sur l'utilisation de la distance de Kullback Leibler. La taille de l'index est de 160 Octets par seconde de musique, permettant donc une indexation compacte.

Dans cet article nous décrivons d'abord l'architecture de l'indexeur CYNDI (CyraNo InDexeur), puis notre technique de segmentation de la bande sonore en phrases basée sur un seuillage automatique, et enfin l'indexation par le contenu de la musique et des effets spéciaux sera décrite.

2 L'architecture de CYNDI

Comme l'illustre la Figure 1, notre indexeur CYNDI comporte les composantes suivantes :

1. Un module de démultiplexage qui sépare le signal sonore d'un programme audiovisuel à l'entrée ;
2. Un module de segmentation en parole ou musique en temps réel avec un délai de 15 secondes ;
3. Un module d'indexation/segmentation de la parole ;
4. Un module d'indexation par le contenu des segments de musique ;
5. Un module de génération des fichiers MPEG7

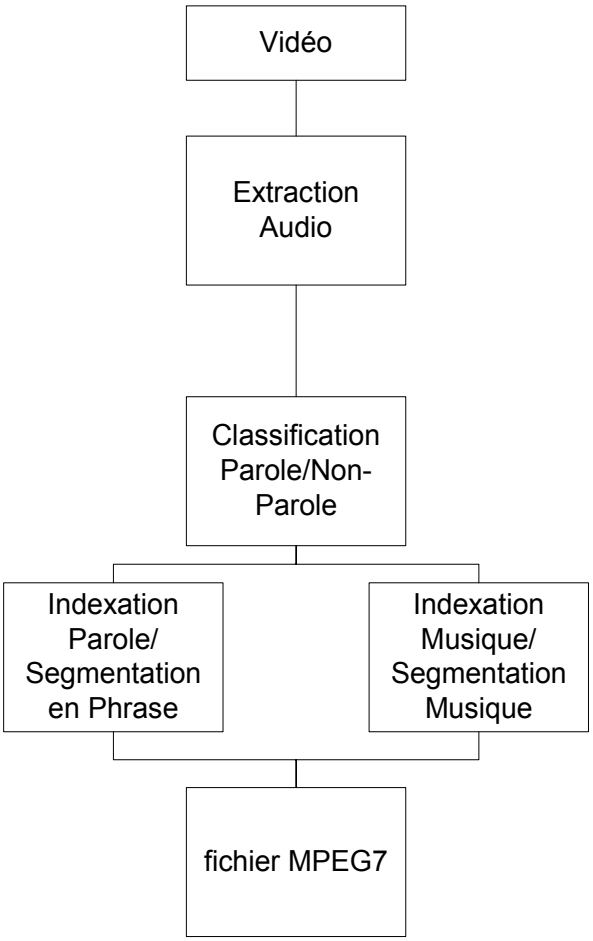

Figure 1 Architecture générale du moteur CYNDI

Dans la suite, nous détaillons les modules 3 et 4, dans la mesure où la segmentation de parole ou de musique d'une bande sonore a déjà fait l'objet de publications antérieures [11].

3 Classification en musique parole

Rappelons que notre moteur d'indexation CYNDI a pour objectif l'indexation automatique en temps réel des flux audio-visuels, tout en permettant un accès ultérieur par le contenu en se basant sur les index générés. Généralement la bande sonore des flux audio-visuels contient des segments de parole, des segments de musique ou effets spéciaux (incluant le bruit) et des segments mixtes contenant plusieurs classes à la fois. Naturellement, les segments de parole doivent être indexés par le texte transcrit de la parole. Contrairement à la parole, la musique et les effets spéciaux ne peuvent pas être efficacement transcrits en une représentation standardisée, même que la transcription automatique de musique ou la reconnaissance des instruments fait l'objet d'un sujet ouvert de recherche. Nous proposons dans notre système une indexation permettant la recherche par requête-exemple du contenu non-parole sans passer par une représentation symbolique.

Clairement une classification du signal sonore en parole/non-parole est cruciale avant l'application d'une méthode ou d'une autre pour l'indexation. Pour cela nous avons développé une technique de classification du signal sonore en des classes sonores [11]. CYNDI utilise cette technique pour la classification du signal sonore en parole/non-parole avant d'appliquer l'indexation.

4 Indexation de la parole

4.1. SEGMENTATION EN PARAGRAPHES PAR UN SEUILLAGE DYNAMIQUE

La segmentation d'une bande de parole en des phrases ou paragraphes poursuit deux objectifs : d'une part elle crée une structure permettant une navigation intelligente et rapide dans un document audiovisuel, et d'autre part elle permet d'améliorer le taux de reconnaissance d'un RAP.
L'approche métrique de segmentation du son a été largement utilisée[1][2][3]. Cette approche consiste à extraire des paramètres du signal sonore dans des fenêtres temporelles glissantes, puis de définir une distance ou mesure de similarité entre les paramètres de deux fenêtres consécutives, une fois cette distance dépasse un certain seuil, une bordure de segment est détectée. L'inconvénient de ces approches est la nécessité de choisir à priori une valeur du seuil qui ne sera pas forcément adaptée aux différents types de documents.
Le critère d'information Bayésienne (BIC Bayesian Information Criteria) a été utilisé par [5][7][9]. Cette approche consiste à chercher les points du signal qui sont le plus probablement des points de changement de

caractéristique acoustique, tel que changement de locuteur, ou de canal.
Une approche utilisable encore est de faire une segmentation suivant une taille fixe de segment, comme dans [4].
Nous adoptons dans notre système l'approche métrique en utilisant la distance de KullBak-Leibler comme mesure de similarité entre deux fenêtres consécutives du signal sonore. Cependant, nous proposons une méthode de calcul automatique du seuil.

Comme nous l'avons signalé, l'inconvénient majeur d'une telle approche est l'utilisation d'une valeur de seuil fixée une fois pour toute empiriquement, alors qu'une telle valeur de seuil peut être fonction du type de document audiovisuel étudié. Nous avons mené une étude statistique sur la variation de longueurs des segments de phrase[1] dans deux heures d'un signal sonore qui est composé d'une heure extraite des journaux télévisés et d'une heure de meetings. Cette étude montre que la variation de longueurs des segments de phrase n'est pas aléatoire, mais suit une distribution presque-Gaussienne, Figure 2.

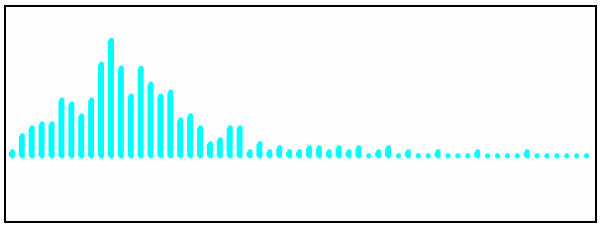

Figure 2 Distribution des longueurs de phrase dans 7200 secondes de parole

Notre idée est d'utiliser ces résultats de distribution de longueurs de phrases pour mieux adapter le seuil.
Une vraisemblance entre la taille d'un segment généré par une technique de segmentation et la distribution statistique des tailles des segments peut être calculée. Ceci permet de choisir entre plusieurs bordures candidates en fixant celles qui maximisent une telle vraisemblance.
Soit un segment de parole de 100 secondes. En appliquant une mesure de similarité sur des fenêtres glissantes sur le signal, une courbe de distance sera générée. Le seuillage de cette courbe permet d'identifier les bordures de segments, Figure 3. Notre idée est de calculer à chaque valeur du seuil une mesure de vraisemblance entre les longueurs de segments générés et la distribution des longueurs de phrases obtenue par notre étude statistique. La courbe dans la Figure 4 nous montre la variation de la vraisemblance par rapport à la valeur du seuil. La valeur du seuil maximisant la vraisemblance sera donc choisie pour aboutir à la segmentation, c'est la valeur la plus probablement optimale. Notons qu'en pratique les segments de tailles inférieures à une seconde ne sont pas pris en compte.
Ceci permet donc de calculer automatiquement pour chaque segment de parole la valeur optimale du seuil. L'inconvénient de l'approche métrique pour la segmentation peut être évité avec cette méthode de calcul automatique du seuil.

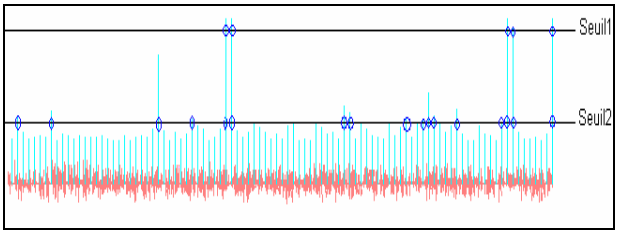

Figure 3 Le principe de seuillage pour obtenir les bordures de segments

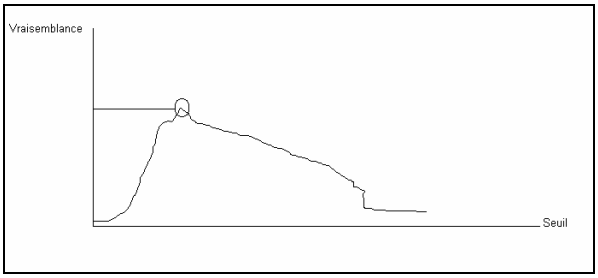

Figure 4 Une courbe montrant la variation de la vraisemblance, entre les longueurs de segments détectés et la distribution de longueurs de phrase, par rapport à la valeur du seuil. La valeur optimal du seuil est la valeur qui maximise la vraisemblance.

La distance entre les paramètres de deux fenêtres consécutives que nous utilisons est la distance de KullBack-Leibler que nous avons étudiée dans un travail antérieur[10]. La distance de KullBack-Leibler (KL) est une mesure de similarité issue de la théorie de l'information. Il s'agit d'une mesure de similarité entre deux variables aléatoires. Dans le cas des distributions Gaussiennes, la distance symétrique de KullBack-Leibler est:

$$KL2(X,Y) = \frac{\sigma_X^2}{\sigma_Y^2} + \frac{\sigma_Y^2}{\sigma_X^2} + (\mu_X - \mu_Y)^2 (\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2})$$

Où $\sigma_X, \sigma_Y, \mu_X, et \mu_Y$ sont respectivement les écarts-types et moyennes de X et Y.
Dans notre cas, X, et Y sont les distributions des vecteurs spectraux issus des fenêtres temporelles consécutives X et Y.

[1] Un changement de phrase inclus un changement de locuteur et du canal ainsi que le changement de phrase classique pour le même locuteur.

4.2. GENERATION DES MOTS CLES

Notre technique de segmentation en phrases permet de découper la bande de parole d'un document audiovisuel en des segments homogènes de longueur moyenne de 15s. Nous appliquons ensuite un moteur de Reconnaissance Automatique de la Parole (RAP) sur les segments ainsi obtenus. Les résultats de reconnaissance de la parole conduisent en général à plus de 50% de Word Error Rate (WER). On constate cependant que dans le type de document audiovisuel que nous considérons, c'est à dire les journaux télévisés ou meetings, les mots les plus importants, qui caractérisent le sujet traité, ont tendance à se répéter. Par exemple, si l'on parle du « nucléaire » dans un meeting, le mot « nucléaire » se répète par la plupart des interlocuteurs, ce qui augmente la probabilité de bonne reconnaissance par le RAP. Nous proposons donc deux contraintes pour l'extraction des mots-clés : 1) le mot se répète plus de 2 fois, 2) le mot contient plus de 2 caractères pour éviter la détection des « de, a, la ,le… » Qui se répètent beaucoup dans le texte.

5 Indexation de la musique et des effets spéciaux

Les efforts en indexation des signaux audio de non-parole sont nettement moins importants que ceux dans le domaine de l'indexation de la parole, en occurrence reconnaissance de la parole. Une approche intéressante d'indexation de la musique ou des effets spéciaux est une indexation qui permet une recherche par requête du même type (musique ou autre). Le système cherche donc dans une base de donnée les segments de musique ou des effets spéciaux qui sont « similaires » à une requête segment de musique ou effets spéciaux [13] [14][15]. Cette approche permet de nombreuses applications dans le domaine de l'accès intelligent aux documents multimédias. A titre d'exemple, un utilisateur peut soumettre une chanson ou la musique de thème d'un programme et le système surveille les sources audio-visuelles pour des documents « similaires » et qui peuvent lui intéresser.
Dans notre moteur d'indexation CYNDI nous utilisons une approche basée sur la distance de KullBack-Leibler pour l'indexation des segments de non-parole.
Comme décrit dans la section précédente, la distance de KL peut être utilisée comme une mesure de similarité entre deux segments de signal sonore, typiquement de durée entre 1 et 2 secondes. Nos expérimentations montrent que cette mesure similarité décrit une similarité acoustique aperçue par les humains.

1. Mesure de similarité

Soit deux segments de non-parole A et B de durée *n* et *m* secondes respectivement. Chaque segments est segmenté en *n* , resp. *m* fenêtres de durée 1 seconde. Les distances KL entre toutes les fenêtres de A et de B constituent une matrice *nxm* de valeurs, la Matrice de Similarité Locale (MSL).

$$MSL_{A,B} = \begin{matrix} KL_{1,1}^{AB} & KL_{2,1}^{AB} & KL_{3,1}^{AB} & \bullet \\ KL_{1,2}^{AB} & KL_{2,2}^{AB} & KL_{3,2}^{AB} & \bullet \\ KL_{1,3}^{AB} & KL_{2,3}^{AB} & KL_{3,3}^{AB} & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{matrix}$$

Nous proposons comme mesure de similarité entre A et B la somme des trois valeurs minimales de la matrice. Notons que plusieurs mesures peuvent être extraites de la matrice. Cependant, dans notre problème d'indexation par le contenu des segments de non-parole, une mesure simple et permettant la recherche par une requête audio est souhaitable, d'où la solution proposée.

2. Indexation

Le fait d'utilisé les distances KL sur des fenêtres de 1 seconde à la base de la mesure de similarité adaptée dans CYNDI pour les segments de non-parole, permet la création des index de ces segments en se basant sur les vecteurs de moyennes et les vecteurs de variance du spectre dans les fenêtres de 1s. Ceci veut dire, que pour chaque seconde du signal, un index de 40 valeurs[2] (20 valeurs de moyenne, 20 valeurs de variance) est nécessaire. Le débit nécessaire pour l'index est donc de 160 octets par seconde.

6 Fichiers MPEG7

CYNDI analyse les flux audio-visuels et génère des fichiers XML compatible avec la norme de description multimédia MPEG7[18]. Les « caractéristiques» extraites automatiquement sont : musique, parole, silence. Pour les segments de musique et de parole le temps de la bordure de segment (phrase pour la parole) sera généré. Les mots-clés constituent l'index des segments de parole et les valeurs de moyennes/variances du spectre de fréquence constituent l'index des segments de musique. Un exemple des fichiers de description générés par CYNDI est présenté dans la Figure 5

[2] Les valeurs extraites du spectre correspondent aux 20 premiers Coefficients Spectraux de MEL, (MEL Frequency Spectral Coefficients)

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Mpeg7>
  - <Video id="jazz.wav">
    - <VideoSegment id="1">
        <feature relevance="0,533333333333333">Music</feature>
      - <MediaTime>
          <MediaTimePoint>0</MediaTimePoint>
          <MediaDuration>15374</MediaDuration>
        </MediaTime>

          <Index>0,00880572333699092;0,687208585441113;0,00258584459515987;0,542035140097141;0,00458
          -
          5;0,110888702329248;0,000861829630594002;0,450806552544236;0,000872545751917642;0,44007669
        <segment_boundary>7000 with prob. : 0,56290205078125</segment_boundary>
      </VideoSegment>
    - <VideoSegment id="2">
        <feature relevance="0,933333333333333">Music</feature>
      - <MediaTime>
          <MediaTimePoint>15374</MediaTimePoint>
          <MediaDuration>14951</MediaDuration>
        </MediaTime>

          <Index>0,013883545761928;1,66960842907429;0,0124474303447641;1,8617320805788;0,03275723720
        <segment_boundary>21374 with prob. : 0,0759896118164063</segment_boundary>
      </VideoSegment>
    - <VideoSegment id="3">
        <feature relevance="0,866666666666667">Music</feature>
      - <MediaTime>
          <MediaTimePoint>30325</MediaTimePoint>
          <MediaDuration>14976</MediaDuration>
        </MediaTime>

          <Index>0,0268398463958874;2,91485302150249;0,0486504402942955;3,8739088922739;0,0746422098
```

Figure 5 Exemple des fichiers MPEG7, sortie de l'indexeur.

7 Implémentation

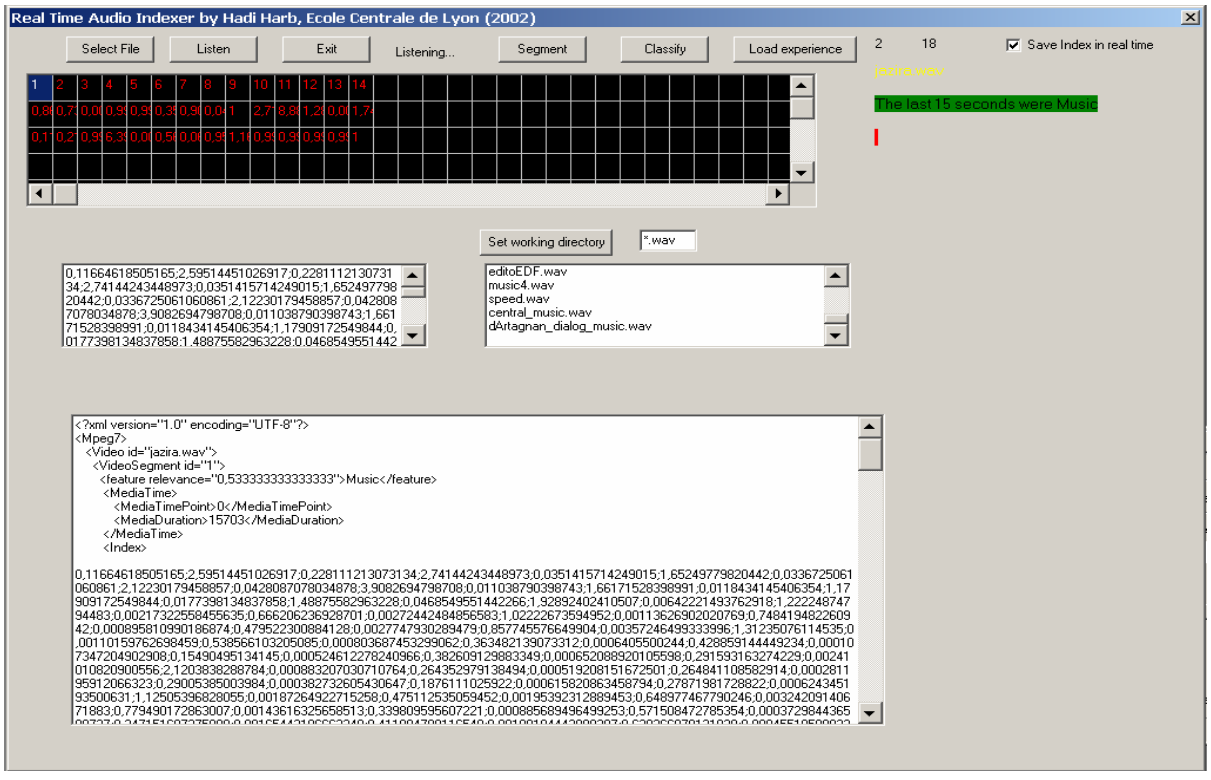

Figure 6 L'interface de notre indexeur CYNDI

Actuellement, CYNDI, notre serveur d'indexation implémenté en C++, Figure 6, analyse et indexe en temps réel, pour une durée de quatre heures par jour, des programmes de la chaîne LCI. Notre base de données contient actuellement plus que quatre cents heures de programmes indexés par mots-clés avec une possibilité de navigation intelligente par saut de « paragraphes » dans les documents. Cette indexation continue permet ultérieurement des expérimentations sur la qualité de recherche de l'information sonore.

8 Conclusion

Dans cet article nous avons présenté CYNDI, notre moteur d'indexation de la bande sonore des documents audio-visuels. CYNDI s'appui sur une classification du signal sonore en musique ou parole avant d'appliquer pour chaque classe une méthode d'indexation spécifique. Pour les segments de parole nous avons proposé une technique de segmentation en phrase en utilisant une nouvelle méthode de calcul automatique du seuil. La transcription de la parole constitue l'index. Pour les segments de non-parole nous proposons une technique d'indexation par le contenu permettant la recherche par requête-exemple.
Les index et les informations sur les bordures de segments sont enregistrés sous forme de fichier XML compatible avec la norme MPEG7.

Références

[1] M. Siegler, U. Jain, B. Ray and R. Stern, Automatic segmentation, classification and clustering of broadcast news audio, Proceedings of the Speech Recognition Workshop, pp 97-99, 1997.

[2] P. C. Woodland, T. Hain, S. Johnson, T. Neisler, A. Tuerk, S. Young, Experiments in Broadcast news transcription, Proc. ICASSP 1998, Seattle, May 1998.

[3] Uday Jain, et al, Recognition of Continuous Broadcast News With Multiple Unknown Speakers and Environments, Proceedings of the Arpa Speech Recognition Workshop 1996.

[4] Ananth Sankar et al, Improved modeling and efficiency for automatic transcription of broadcast news ,Elsevier, Speech Communication, 2001

[5] P. Beyerlein et al, Large vocabulary continuous speech recognition of broadcast news - The Philips/RWTH approach, Elsevier, Speech Communication, 2001

[6] Jean-Luc Gauvain, Lori Lamel, Gilles Adda, The LIMSI broadcast news transcription system ,Elsevier, Speech Communication, 2001

[7] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in DARPA speech recognition workshop, 1998

[8] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, Strategies for automatic segmentation of audio data, Proc. icassp2000 ISTANBUL pp 1423-6, 2000

[9] Mauro Cettolo and Marcello Federico, Model selection criteria for acoustic segmentation, in Proc. of the ISCA ITRW ASR2000 Automatic Speech Recognition, Paris, France, 2000, pp. 221--227.

[10] Hadi Harb, Liming Chen, Jean-Yves Auloge segmentation du son en se basant sur la distance de KulBack-Leibler, pp 63-68 CORESA 01 Dijon France, 2001.

[11] Hadi Harb, Liming Chen, "Technique de classification du signal sonore en des classes sonores", pending patent nf 02 08 548, Juillet 2002

[12] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley Series in Telecommunications. John Wiley and Sons, 1991

[13] Jonathan T. Foote, Multimedia Storage and Archiving Systems II, Proc. of SPIE, Vol. 3229, pp. 138-147, 1997

[14] Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search and retrieval of audio", IEEE Multimedia Magazine, vol. 3, no. 3, pp.27-36, 1996

[15] Z. Liu and Q. Huang, "Content-based indexing and retrieval by example in audio," in Proceedings of IEEE ICME 2000, 2000.

[16] Beth Logan and Ariel Salomon, A MUSIC SIMILARITY FUNCTION BASED ON SIGNAL ANALYSIS, in Proceedings of IEEE ICME 2001, 2001.

[17] S. Z. Li, "Content-based classification and retrieval of audio using the nearest feature line method", IEEE Trans. on Speech and Audio Processing, Sep., 2000.

[18] ISO/IEC draft MPEG7 Audio specifications, http://www.darmstadt.gmd.de/mobile/MPEG7/Documents.html, N3802.