International Journal on Digital Libraries (2006) 6(1): 70–81
DOI 10.1007/s00799-005-0120-5

REGULAR PAPER

Hadi Harb · Liming Chen

Audio-based description and structuring of videos

Published online: 23 February 2006

Abstract Enabling a rapid on-the-fly view of the content of a movie requires segmenting the movie and describing the segments in a user-compatible manner. The difficulty resides in extracting relevant semantic information from the audiovisual signal, both for the segmentation and the description. We introduce in this paper audio scenes and chapters in movies and present an algorithm for automatically segmenting a video based on the audio stream only. Audio scenes and chapters are defined as the equivalent of shots and scenes in the visual domain. A tree-like audio-based structure of a video is proposed. A chapter is then classified into different chapter categories. The automatic solution to audio scene and chapter segmentation and classification is evaluated on manually segmented and classified videos

Keywords Video segmentation · Audio classification · Audio scenes · Piecewise Gaussian Model

1 Introduction

Structuring a video based on its semantic content is of increased relevance to allow intelligent video navigators, skimmers and search engines with the increase in the use of digital video archives. Video structuring is the problem of video segmentation into semantic units, also called story units, and the creation of a storyboard similar to the table of contents in a book. Automatically generating a storyboard of a video will probably enhance the experience of video retrieval both for consumer and professional applications. The objective of video structuring and description is to permit a user to have an on-the-fly idea of the content of a video or a scene.

In this paper, the term semantics is defined as a mid or high level interpretation of low-level signal features. Examples of mid-level interpretation include for instance, speech, music, cars, trees.... Higher level understanding of these mid-level features, such as the classification of video content into "calm dialog", "battle", "car pursuit", etc. is a real challenge for the content-based multimedia indexing community. This paper tries to use low- and mid-level audio features for content-based video structuring and description.

Starting from basic units such as shots, video structuring consists of grouping these units into scenes that are semantically homogenous and grouping scenes into chapters and iteratively continuing the grouping till covering the entire document. For example, several shots of one telephone conversation are considered as one scene. The underlying action is the telephone conversation. Several scenes telling one underlying story may be grouped together in one chapter. This kind of grouping is a semantic grouping based on a high level of understanding of the content. Although video segmentation into shots or the grouping of similar shots into scenes can be considered as a mature domain, semantic scene and chapter segmentation seems a harder problem. The main reason for the difficulty of scene, or story unit, determination is the high level of semantics engaged in the process, making even a manual segmentation of a video into scenes and chapters subjective and difficult to some extent.

In this paper, we consider structuring a video based on the audio stream only. We restrict our analysis to movies since they offer a major component in the entertainment market especially with future applications such as interactive TV or Video On Demand services.

The main motivation behind our analysis of the audio stream for video structuring is a preliminary experiment that we have conducted on the ability of human subjects to detect scene changes when listening to a movie with no information on the visual stream. We have observed that human subjects were able to structure a movie into semantic units even when they do not understand the language of the movie. Human subjects generally base their judgments on the combination of the acoustical environment and the changes in the mood, described by the mood of speech and music. This ability of understanding the semantic structure

H. Harb (✉) · L. Chen
Département Maths-Info, LIRIS Lab., CNRS FRE 2672, Ecole Centrale de Lyon, France
E-mail: {Hadi.Harb, Liming.Chen}@ec-lyon.fr

of a movie based on the audio stream only is probably due to the extensive use of the music and audio effects by the movie makers in order to convey semantic information to the spectators.

Moreover, the use of audio information for video segmentation offers the advantage of being computationally less demanding and having fewer dimensions than the visual information.

As a movie is classically structured into frames, shots, scenes, and chapters, we introduce in this paper an audio-only structure into audio scenes, and audio chapters at different levels of abstraction. An audio scene is defined as the time instants of a video containing a homogenous acoustical environment and a homogenous semantic environment, such as calm speech or loud music. The same criterion is used to group several consecutive audio scenes into audio chapters.

A table of contents of a movie is therefore created. The algorithm presented in this paper for the segmentation of the audio stream is based on a combination of spectral and semantic dissimilarities. We also present a grouping algorithm for merging consecutive audio scenes into scenes, or chapters, having a higher level of abstraction. Each chapter is moreover classified into different chapter categories.

2 Related work

Video scene segmentation has gained an important effort from the research community. In the majority of cases, the segmentation is based on an analysis of the visual stream of a video. Generally the video is first segmented into shots, a domain that can be considered as mature at present with high accuracy algorithms and standard methodologies for evaluation [1, 2]. Several shots are then grouped together on the basis of similarities in their color histograms, objects, motion, rhythm, and so on [3–6].

The audio stream has been considered in some cases as a complement to the visual stream. Simple algorithms for audio analysis and segmentation were generally used.

Another family of algorithms for video segmentation is the one considering an audio-only segmentation. Few papers in the literature fall in this category, to which the work presented here belongs.

The basic approach generally used for audio-based video segmentation can be considered as a blind segmentation. It consists of defining a similarity measure between neighboring time windows based on signal features such as spectral or cepstral ones. Peaks in the similarity measure correspond to potential scene boundaries.

Sundaram presents in [7] an audio scene segmentation technique based on low-level audio signal features, such as cepstral and energy features. The technique is based on a correlation measure between the envelope of the audio features in an attention time window (16 s) and the envelope of the features in a memory time window (17 s). This technique can actually be considered as a technique based on a measure of similarity between consecutive time windows, attention and memory windows. Cao et al. [8] demonstrate experimentally that a distance between spectral distributions in two consecutive time windows is more convenient than the correlation measure presented by Sundaram.

Minami et al. in [9] try to use a simple speech/music classification in videos in order to enable novel video browsers. The authors use information about the peaks in the spectrogram for speech and music classification. A comprehensive description of the significance of the audio stream in a video is presented in the paper; the authors base their analysis on video production rules. A video browser application is developed based on the speech/music classification.

Pfeiffer in [10] presents an algorithm for video segmentation based on audio analysis. The author uses spectral vectors as audio signal features. The Euclidian distance between vectors from present and a mean of an exponential prevision of the vectors from the past is used as a measure of similarity to detect scene boundaries.

In Ref. [11], shots are merged together using an audio-based similarity measure. The audio similarity measure is based on a Euclidian distance of the means and variances of low-level audio features such as Energy, Spectral centroid, and Spectral flux.

Atalan in [12] combines basic speech/music/silence detection with face detection in order to detect dialog scenes using a Hidden Markov Model. The speech/music/silence detection uses energy thresholding and frequency analysis.

A simple speech/music/silence detection algorithm is used in [13] in order to improve shot and scene detection in videos.

From this brief review of the state-of-the art in audio-assisted video scene segmentation, we can conclude that the proposed algorithms are more likely to be compared to shot segmentation in the visual domain. The segmentation process is generally based on low level acoustic features and sliding time windows. Whereas the process of generating scenes from shots requires a higher level of understanding of the video content, such as objects, rhythm and other semantic information. The audio segmentation presented in the literature, since it is based on low-level acoustic features and generally with little or no knowledge about the semantic content, can be considered as a blind segmentation. We argue that including semantic information to the segmentation process of the audio stream will permit a segmentation level more similar to that of scenes than to shots. Moreover, this semantic information may be used to describe the content of a scene.

Consider for instance a dialog scene between two persons, one from indoor and the other from outdoor. The acoustical environment of the two persons will be probably different. The mismatch in the acoustical environment will be translated by different characteristics if analysed in the low-level feature space, such as the cepstral or spectral feature space. Figure 1 illustrates an example from an English movie “Bridget Jones Diary”; it consists of a dialog scene

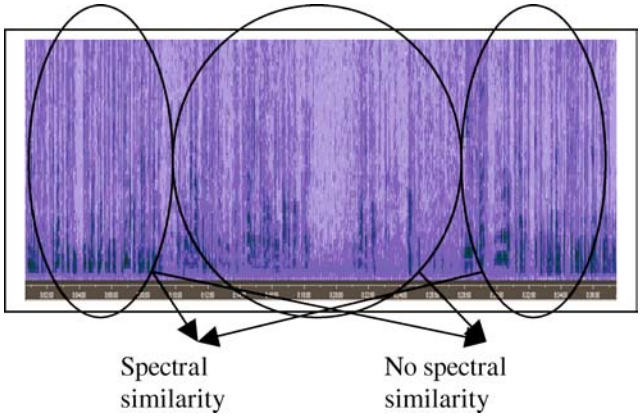

Fig. 1 Spectral similarity does not always convey a semantic similarity. An example where a dialog scene can be segmented into three different scenes if a spectral similarity (e.g. Euclidian distance between the means of spectral vectors in consecutive time windows of n seconds) is used to segment the signal

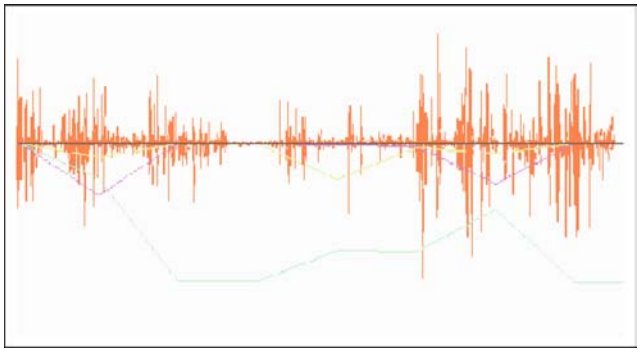

Fig. 2 The continuity of the content, speech, provides a means to eliminate false scene boundaries provided by a spectral analysis

where the two persons have each a different environment, an office and a hall for instance. As we can see on the spectrogram, the different environments are translated into different spectral characteristics. If we apply a similarity measure on the low-level spectral features, a peak will be obtained suggesting a false scene boundary.[1] Conversely, the content, speech for instance, has a continuity which does not lead to a scene boundary. The continuity of the content is illustrated in Fig. 2 which shows the output of a speech/music/noise classifier.

Consider another example where music is included at a given time instant. If the energy of the musical content is relatively weak, a similarity measure based on the low-level features cannot detect a scene boundary. Knowing that the inclusion of music is generally used by movie makers to translate an important change in the underlying story implies a missed detection of a semantic scene boundary in the case of a blind segmentation.

[1] An example of a similarity measure that can be used is an Euclidian distance applied to the mean of the spectral vectors in two overlapped time windows of 1 s.

The content continuity constitutes a complementary source of information, which, if combined to the acoustical continuity, may improve the segmentation process.

To conclude, few works studied the problem of video scene segmentation based on the audio stream. Furthermore, these works consider the segmentation of the audio stream based on low-level signal features with no or little knowledge about the content. In this paper, we propose to combine acoustic and content information for audio scene segmentation and description.

3 Audio scenes

An audio scene is defined as the time instants of a video containing a homogenous acoustical environment and a homogenous semantic environment, such as calm speech or loud music. By following this definition, the acoustical homogeneity leads to a blind segmentation while the content or semantic homogeneity leads to a content-aware segmentation. The acoustical segmentation is equivalent to a raw segmentation applied directly to low-level features; while the semantic segmentation is obtained after an interpretation of the low-level features. Their combination defines audio scenes. The mean duration of an audio scene in a typical movie is about 20 s. An audio scene boundary is not always confirmed by a visual scene boundary; a speech in a dialog can continue after a shot boundary for example. It is important to notice that a change in the semantic content is not always transcribed by a change in the visual features usually used by researchers, such as colour or objects. On the other hand, when the movie-maker intends to signal an important semantic action where no change in the decoration or place is observed, the music is generally used as the basic tool to translate the intention. Consequently, the audio stream conveys important semantic information needed for the understanding of the movie.

Several consecutive audio scenes grouped together on the basis of the similarity between their audio content constitute an audio chapter. Based on this definition, several indoor-audio scenes are grouped together while outdoor scenes will constitute a different audio chapter since the audio characteristics of indoor-audio scenes will probably be similar while being dissimilar from that of outdoor-audio scenes. A typical duration of an audio chapter is about 100 s, [7, 8].Therefore, a full-length movie will contain about 70 chapters.

3.1 Scene description

The video documents we are considering in this work are those containing a variety of sound classes, movies for instance. That is, we do not consider documents containing speech-only audio content, such as meetings.

The audio content of a movie conveys several semantic classes. These classes include, calm dialog, active dialog, neutral dialog, natural action, special effects action, emotion (happiness, sadness) and fear. The extraction of semantics

Table 1 Audio scene categories in terms of speech, music, and noise

	Music	Speech	Noise
Neutral dialog	+	+++++	+
Active dialog	+	+++	++
Calm dialog	++	+++	+
Emotion (happiness, sadness, etc.)	+++++	+	+
Fear	++++	+	++
Action	+	+	+++++
Special effects action	++	+	+++

from the audio signal can be based on a general audio classifier trained on the selected classes. Unfortunately, sufficient training data for the selected semantic classes or concepts is difficult to obtain due to the great variety of audio signal in each class. Nevertheless, intuitively, each of the audio semantic classes can be partially described by a combination of speech, music, and noise. To verify this hypothesis, the first author manually segmented and classified the scenes of two movies, Gladiator and Taxi2, into music, speech, noise, music+speech, music+noise, and speech+noise. For each scene, the percentage of speech, music and noise is obtained. In this case, the percentages do not sum to 100 since in the cases of coexistence of two classes, the time duration is counted two times.

A synthesis of the manually-obtained descriptions in terms of music, speech, noise classes is presented in Table 1. In this table a "neutral dialog" corresponds to a dialog or monolog where only the information is presented such as the case of a news program presenter. An "active dialog" corresponds to scenes containing shouting dialogs or monologs, such as before-fight aggressive dialog. A "calm dialog" is a scene where the dialog or monolog is slow, such as a romantic dialog. "No voice emotion" means scenes where no voice exists but an emotion is conveyed, such as a death scene. "Fear" relates to scenes conveying a disturbance without being sadness or happiness. "Natural action" corresponds to scenes where the action is natural without the intervention of a human artistic creation. "Special effects action" relates to scenes where the music or other special effects are added to convey the intention, such as a battle scene. We distinguish between natural action and special effects actions since each category may convey a different level of action. Generally special effects action is more intensive than the natural action due to the musical component.

3.2 Towards automatic scene description

The preliminary study presented previously encourages a rule-based approach applied on speech, music and noise percentages in order to describe the semantic content of an audio scene. Once the percentages of speech, music and noise components are obtained, a set of rules can be applied describing therefore the content of a scene.

In this work, the percentages of speech/music/noise components are obtained using a Piecewise Gaussian Model–Multi Layer Perceptron (PGM–MLP) audio classifier.

3.2.1 An overview of the PGM–MLP audio classifier

The PGM–MLP audio classifier is inspired by some aspects of the human perception of audio classes [14]. It is fair to suppose that the perception of a stimulus is strongly related to its correlation with the past memory. The context effect is a known effect in the human speech recognition domain. We model the context effect by an auditory memory and we suppose that the classification of a stimulus at time instant t is based on the status of the memory at time instant t. The auditory memory is supposed to be a Gaussian distribution of the spectrum in the past time window, called the Integration Time Window (ITW). The auditory memory is therefore modelled by one Gaussian distribution for each frequency band. The frequency bands are chosen according to the MEL psychoacoustic scale in order to model the frequency resolution of the human ear. The auditory memory model is updated continuously by a new acoustic observation and hence, by new spectral features. For the sake of simplicity, we suppose that the duration of the memory model is constant, which means that the ITW is constant. Also, the Gaussian distributions are described by their mean and diagonal covariance.

$$\vec{\mu}^{(t+1)} = (1-\varepsilon)\vec{\mu}^{(t)} + \varepsilon\vec{X}^{(t+1)} \tag{1}$$

$$\vec{\sigma}^{(t+1)} = (1-\varepsilon)\vec{\sigma}^{(t)} + \varepsilon(\vec{\mu}^{(t+1)} - \vec{X}^{(t+1)})(\vec{\mu}^{(t+1)} - \vec{X}^{(t+1)})^T \tag{2}$$

$\vec{X}^{(t)}$ is the spectral vector, the Mel Frequency Spectral Coefficient (MFSC) vector for instance [15], $\vec{\sigma}^{(t)}$ and $\vec{\mu}^{(t)}$ are the variance and the mean, respectively, of the short-term spectral vectors at the time (t). ε is a decay parameter that governs the speed of forgetting the past. A simple value of ε is a constant over time that equals $1/T$, where T is the number of short term spectral vectors contained in the ITW window.

On the basis of this interpretation, the Piecewise Gaussian Modelling consists of the Gaussian parameters of the short-term memory model. The PGM features are therefore the normalised mean and variance of the Gaussian distributions for each ITW window.

We can summarize the PGM feature extraction by the following [16]:

1. The computation of the MFSC features with 30 ms Hamming window and 20 ms overlap. 20 Mel scaled filters are used and hence one vector containing 20 elements is obtained each 10 ms.
2. Grouping MFSC vectors in ITW windows, estimating the Gaussian parameters of each ITW window. The Gaussian parameters are the mean and variance vectors and constitute 2 vectors of 20 elements each.
3. Normalising the mean values by their respective maximum and normalising the variance values by their respective maximum.

The duration of the ITW windows is a function of the audio classes. Typically we use 4 s as a standard duration in the case of Audio Scene Segmentation.

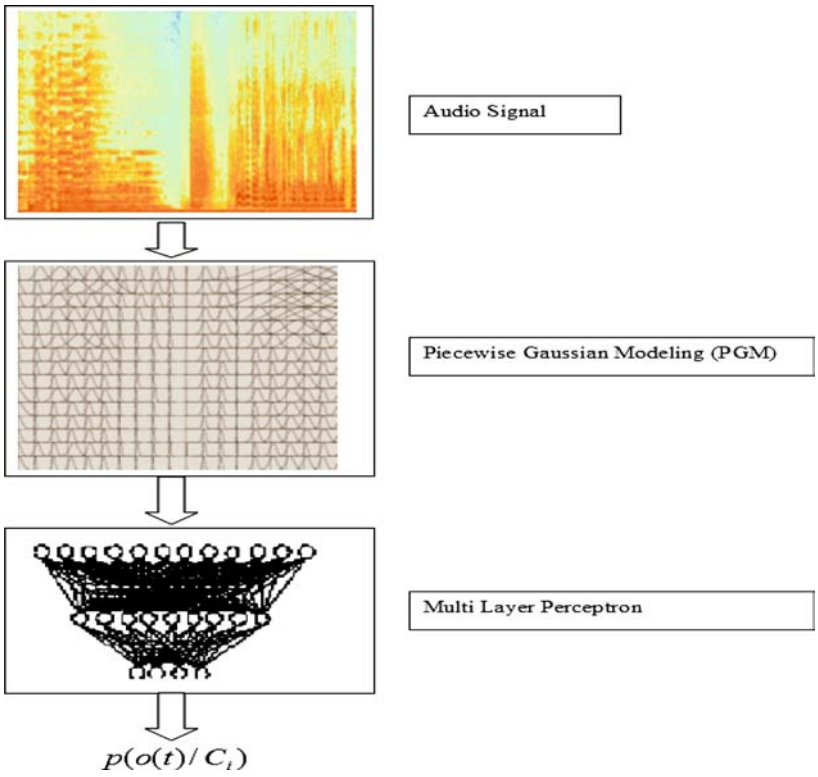

Fig. 3 The PGM–MLP audio classifier's architecture. Each column in the middle section corresponds to one feature vector (the Gaussian parameters in each frequency band)

The PGM features are then coupled to a Multi Layer Perceptron (MLP) trained using the error back-propagation algorithm. The MLP has 40 input neurons corresponding to the 20 mean values and the 20 variance values obtained from each ITW window. The MLP has 1 hidden layer with 100 hidden nodes, 3 output neurons in the case of speech/music/noise classification, and is fully connected. The MLP estimates, after a training phase using the gradient-descent algorithm, the probability of the audio classes given the PGM features of the ITW window. The architecture of a PGM–MLP classifier is shown in Fig. 3.

3.2.2 Rule-based audio scene description

The rules, obtained from Table 1, that have been used in order to describe the content of a scene/chapter are the following:

S>>N>M	Active dialog (AD)
S>>M = N	Neutral dialog (ND)
N∼>S>M	Natural action (NA)
M>N∼>S	Emotion no voice (E)
M>S∼>N	Emotion no voice (E)
N>M>>S	Special effects action (SA)
N>M>S	Special effects action (SA)
S∼>M>N	Calm dialog (CD)
S>M>N	Calm dialog (CD)
M>N>>S	Fear (F)
M>N>S	Fear (F)

where x>>y means x>2y, x∼>y means y<x< 1.05y, and x>y means x>1.05y.

These rules are basic and heuristic rules. It is also possible to obtain similar rules by using a Decision Tree algorithm or other types of classifiers which may provide better results.

The presented rules were self-validated using a speech/music/noise classifier on the same data used for their manual elaboration. That is, a speech/music/noise classifier, the same as the one described in Sect. 3.2.1., was used to classify the audio stream of the first 1600 s of the movie "Gladiator". The classifier was trained on 240 s of training data (80 s for each of the audio classes). The training data was obtained from the "Gladiator" movie excluding the test data. The ITW duration was set to 4 s. Therefore, the MLP was trained on 60 training samples. The rules were generated and the automatically obtained audio classes were compared to the manually obtained ones for the manually segmented chapters. In this experiment "Neutral dialog" and "Fear" were not included due to insufficient data in the test segment. Table 2 shows the confusion matrix between the automatically obtained labels and the manually obtained ones of the chapters. Globally 67% of chapters were correctly classified. Notice that the errors of the audio classifier combined with the over-simplification of rules affect the results. Nevertheless, these results suggest that a mapping can be done from basic audio classes to semantic ones via simple rules.

Table 2 Confusion matrix between the automatically obtained scene categories and the manually obtained ones

	Real				
System	A.D.	C.D.	E.	S.A.	N.A.
A.D	2				
C.D	1	2			
E	1		2		1
S.A				2	
N.A				2	2
Total	4	2	2	4	3

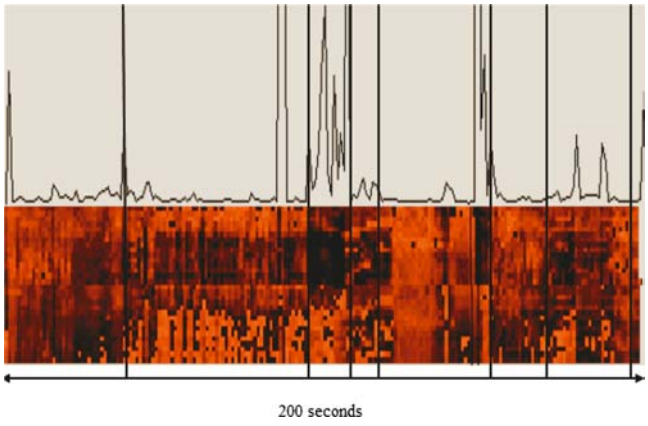

Fig. 4 The Acoustic Dissimilarity Measure values and the real scene boundaries (*vertical lines*)

4 Audio scene segmentation

Audio-scene segmentation is the first step required to delimit scene boundaries.

In this work, we make use of a combination of two dissimilarity measures: the Acoustic Dissimilarity Measure (ADM) and the Semantic Dissimilarity Measure (SDM).

4.1 Acoustic Dissimilarity Measure (ADM)

The change in the place or the change in the sources is acoustically transcribed by changes in the spectral environment. That is, the frequency spectra of office noise have generally different characteristics than that of a street noise since the characteristics of the acoustical sources are different. This can be observed while observing the spectrogram of the audio stream of a movie.

We use the Kullback-Leibler distance between consecutive time windows to translate the changes in the acoustical environment in a movie since it has been demonstrated suitable for the estimation of spectral audio similarity [17]. Other approaches may be used for this task of acoustic segmentation such as the Bayesian Information Criterion (BIC) approach [18]. The Kullback-Leibler distance is however more convenient when using the PGM–MLP classifier due to the fact that they use the same features.

The Kullback-Leibler (KL) distance originates from the information theory. It is a distance between two random variables. The original KL distance does not have the properties of a distance, but the symmetric KL is a distance [19]. In the case of Gaussian distribution of the random variables the symmetric KL distance is computed by:

$$KL2(t) = \frac{\sigma_t^2}{\sigma_{t+1}^2} + \frac{\sigma_{t+1}^2}{\sigma_t^2} + (\mu_t - \mu_{t+1})^2 \left(\frac{1}{\sigma_t^2} + \frac{1}{\sigma_{t+1}^2} \right) \quad (3)$$

μ_t and σ_t^2 are respectively the mean and the variance of the spectral variables in the t time window. In the case of mean and variance vectors the Acoustic Dissimilarity Measure is computed as:

$$ADM(t) = KL2(t) = \frac{\vec{\sigma}_t \vec{\sigma}_t^T}{\vec{\sigma}_{t+1} \vec{\sigma}_{t+1}^T} + \frac{\vec{\sigma}_{t+1} \vec{\sigma}_{t+1}^T}{\vec{\sigma}_t \vec{\sigma}_t^T} + (\vec{\mu}_t - \vec{\mu}_{t+1})(\vec{\mu}_t - \vec{\mu}_{t+1})^T \left(\frac{1}{\vec{\sigma}_t \vec{\sigma}_t^T} + \frac{1}{\vec{\sigma}_{t+1} \vec{\sigma}_{t+1}^T} \right) \quad (4)$$

where $\vec{\mu}^T$ and $\vec{\sigma}^T$ are respectively the transpositions of $\vec{\mu}$ and $\vec{\sigma}$.

Since very short-term changes in the spectrum, such as the silence between two words, are not relevant for scene changes, we consider relatively long-term time windows, that we call Integration Time Window ITW, for the estimation of the KL2 distance. The typical duration of the ITW window is 4 s. Figure 4 shows an example of the ADM values for a segment of 200 s from Italian movie "Malena". As we can see on the figure, the ADM measure can predict a portion of the scene boundaries although it can be considered as an optimal predictor.

4.2 Semantic Dissimilarity Measure (SDM)

The time instant of a change in the content is probably a scene boundary. For instance, a transition from a calm dialog to a violent one is generally a semantic scene boundary. Therefore, estimating the novelty of the audio content at each time instant is one important feature for scene segmentation.

Let $P_i^{(t)} = p(C_i \,|o(t))$ be the probability of the audio class C_i, given the acoustic observation, or feature vector $o(t)$, at the time instant t, $i = 1, \ldots, N$, N represents the number of classes.

The SDM measure is computed as follows:

$$SDM(t) = \sqrt{\sum_{i=1..N} \left(P_i^{(t)} - P_i^{(t-1)} \right)^2} \quad (5)$$

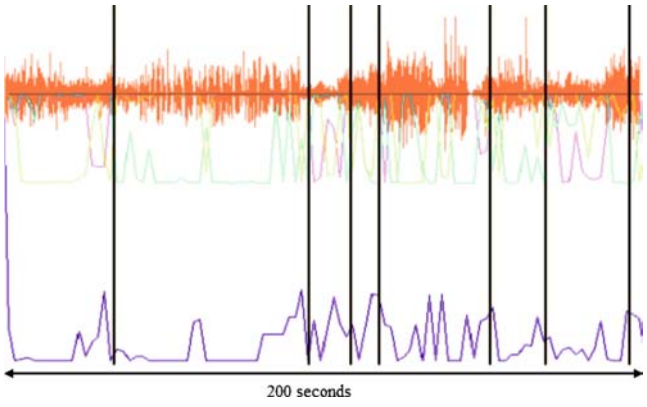

Fig. 5 The Semantic Dissimilarity Measure values and the real scene boundaries (*vertical lines*)

In this work we consider three basic audio classes, speech, music and noise due to the availability of the training data for these audio classes. We make use of the PGM–MLP general audio classifier for the estimation of the probabilities of each of the three classes, given the acoustic observation. Figure 5 shows the same example of Fig. 4 and the corresponding SMD measure. As we can see, false and/or missed scene detection from one similarity measure can be corrected by the other similarity measure. We state that combining the acoustic and the semantic information and hence combining ADM and SDM is advantageous; none of the measures is sufficient alone for an effective segmentation.

4.3 Audio Scene Segmentation based on the ADM and the SDM

The ADM values for a given video document are in a different interval than that of the SDM. Therefore, we normalise the ADM values in order to have the same mean as that of the SDM values for a given document. The novelty in the audio content at each time instant is given as a weighted sum of the normalised ADM values with the SDM values.

The novelty in the audio content can be obtained as follows:

$$\mathrm{K}(t) = \alpha.SDM(t) + \beta.ADM(t) \tag{6}$$

α and β are two suitable weights.

A correlation analysis between the values of the SDM and ADM on the first 1600 s from six movies was carried out. The correlation is 0.0022, suggesting that no direct relation exists between SDM and ADM, and hence the combination of the two measures provides a combination of relatively independent sources of information.

4.3.1 Choosing weights

The respective weights used in order to combine the ADM and the SDM measures need to be set in accordance to the relative importance of each of the measures for the scene segmentation task. In this work, the weights are chosen after a training phase on the two movies used to train the audio classifier. Given a manual segmentation into scenes, the correlation between the true scene boundaries and the values of the dissimilarity measures gives a fair estimate of relative importance for each measure.

The training data corresponds to the first 1600 s from the two movies "Gladiator" and "Taxi2". The scene boundaries are used to provide the *TB(t)* function. The values of the function are obtained as follows:

$$TB(t) = \begin{cases} 1 & \text{if } t = \text{sceneBoundary} \\ 1 & \text{if } t-1 = \text{sceneBoundary} \\ 1 & \text{if } t+1 = \text{sceneBoundary} \\ 0 & \text{Otherwise} \end{cases}$$

The time t is quantized with a step equals to ITW/2.

α and β are obtained as follows

$$\alpha = \frac{\sum_{t=0}^{2M} TB(t).SDM(t)}{\sum_{t=0}^{2M} SDM(t)} \tag{7}$$

$$\beta = \frac{\sum_{t=0}^{2M} TB(t).ADM(t)}{\sum_{t=0}^{2M} ADM(t)} \tag{8}$$

where M corresponds to the number of ITW windows in the training data.

4.3.2 Choosing threshold

If a static threshold is used to detect the scene boundaries we risk obtaining very short scenes and very long scenes if higher values are concentrated in one portion of the document (Fig. 6). As a consequence, we detect the scene boundaries as local maxima in a sliding window having the same duration of the desired scene duration (Fig. 7). Therefore, no threshold is needed to be set for each video document. Instead, the duration of a scene is to be set as a parameter. We use 20 s as a standard value for the scene duration.

5 From audio scenes to audio chapters

The content of an audio scene is mainly described by the semantic and the acoustical contents. The semantic content indicates whether a scene is mainly a dialog, violent, calm, sensitive, or an action, or music. On the other hand, the acoustical content describes the scene in a blind manner where only information about the power of the signal in each

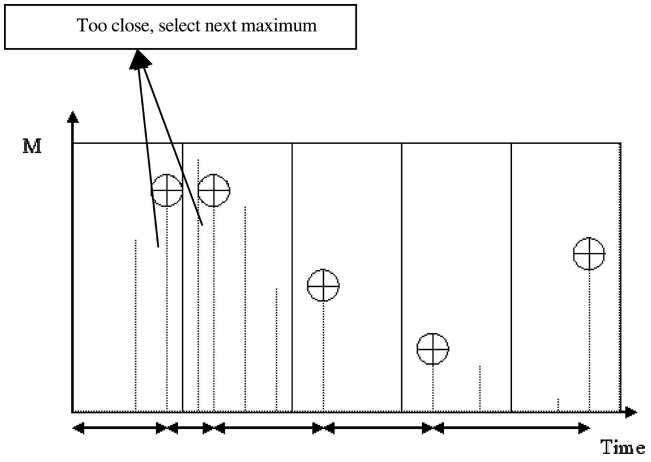

Fig. 6 Local maxima are selected as scene boundaries. When two local maxima are to close (less than 10 s between them) the next maximum is chosen as the scene boundary

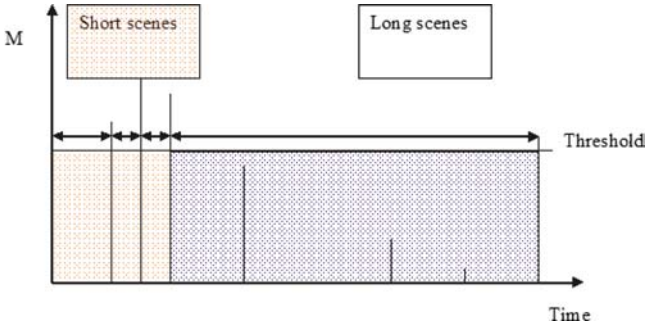

Fig. 7 Static thresholds will probably provide very short scenes and very long scenes depending on the nature of the document

frequency band is gathered. A suitable measure of similarity between two audio scenes will combine the two sources of information, the semantic and the acoustical one. That is, two scenes containing a violent dialog are "similar" if the semantic content is used, while two scenes shot in the street are "similar" if the acoustical content is used.

The semantic content of a scene is described by a vector containing the mean and the standard deviation of the probabilities of each of the audio classes, speech, music and noise for instance:

$$\vec{V}s = \begin{bmatrix} meanP_N \\ meanP_M \\ meanP_S \\ stdvP_N \\ stdvP_M \\ stdvP_S \end{bmatrix}$$

where $meanP_i$ and $stdvP_i$ are the mean and the standard deviation of the probabilities of audio class i in the scene.

We use the Euclidian distance between the global semantic vectors of two scenes in order to estimate the semantic similarity.

The acoustical content of a scene is described by the mean and the standard deviation of the acoustic vectors contained in a scene.

$$\vec{V}a = \begin{bmatrix} meanX1 \\ meanX2 \\ . \\ . \\ stdvX1 \\ . \\ . \end{bmatrix}$$

Where $meanX_i$ and $stdvX_i$ are the mean and the standard deviation of the spectral components of frequency band i of the audio signal in the scene.

The Kullback-Leibler distance is used to estimate the acoustic similarity between two scenes on the basis of the mean and the variance of the acoustic vectors.

The similarity measure between two consecutive scenes, i *and* j, is a linear combination of the semantic similarity and the acoustic similarity.

$$M(i,j) = \alpha.d\left(\vec{V}_S^{\,i}, \vec{V}_S^{\,j}\right) + \beta.d\left(\vec{V}_a^{\,i}, \vec{V}_a^{\,j}\right) \quad (9)$$

The local maxima in a sliding window are picked as chapter candidates. The duration of the sliding window is a parameter fixed by the user which affects the mean duration of a chapter. We use 80 s as a standard duration in our experiments.

The process of chapter generation can be iterated given scenes from a lower level of abstraction. Therefore, we can create a tree representing a movie where we can start the navigation from higher levels of abstraction to lower levels. Each unit, scene or chapter, is further described by the percentages of the speech, music and noise classes (Fig. 8). The user can have an on-the-fly idea of the content of a scene when looking at these percentages. This tree constitutes the table of contents, or the storyboard of a movie. A representative image can be included for the description of a unit in the tree. The audio stream is further classified using the speech/music/noise percentages and the rules from Sect. 3.2.2 into different categories, Natural Action, Special Effects Action, Dialog, Active Dialog, Calm Dialog, Emotion

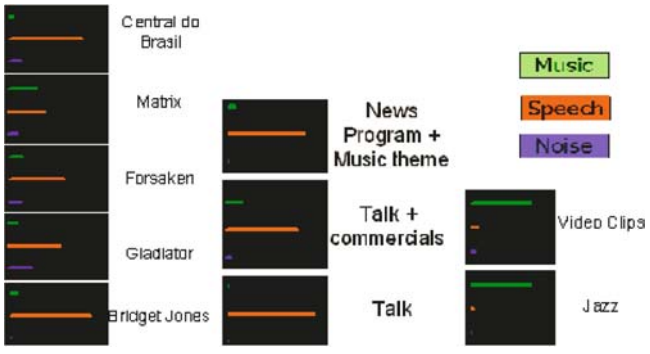

Fig. 8 Examples of the speech/music/noise probabilities of different types of video documents

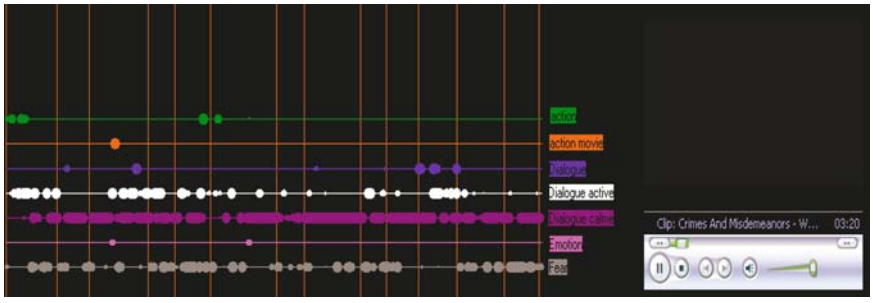

Fig. 9 The interface of the system, chapter boundaries at a specific level of abstraction and the description of the content (in terms of the different scene categories) are shown to the user in order to enable easy navigation and visualization of the document

and Fear. The scene categories and the scene boundaries at a given level of abstraction can be visualised via a graphical interface such as the one shown in Fig. 9.

6 Evaluation

The proposed audio scene segmentation algorithm and the rule-based classification algorithm were evaluated on a dataset constituted by the first 1600 s of four movies.

6.1 Segmentation

The manual segmentation of a movie into chapters, and semantically homogenous scenes includes an important part of subjectivity making the results no more than a partial estimation of the performance. However, we performed a manual segmentation into chapters of four movies and we evaluated the system in terms of precision and recall ratios. The recall and precision ratios are computed as follows:

$$\text{recall} = \frac{\text{manual boundaries detected by the system}}{\text{total manual boundaries}} \tag{10}$$

$$\text{precision} = \frac{\text{manual boundaries detected by the system}}{\text{total boundaries detected by the system}} \tag{11}$$

The first 1600 s from the "casino", "Artificial intelligence", "Training day" and "Life is beautiful" movies were considered for the experiment. The PGM–MLP audio classifier was trained on 80 s of speech, 80 s of music and 80 s of noise from "gladiator" and "Taxi 2" movies. The ITW duration was set to 4 s and therefore the MLP was trained on 60 training samples. The MLP has 40 input neurons, 1 hidden layer with 100 neurons, and 3 output neurons. The MLP is fully connected. The performance was measured as the optimal performance in three runs. The mean audio scene and chapter durations were set to 20 s and 80 s, respectively.

The manually obtained chapters were based on the combination of the visual and auditory modalities. It is important to compare the automatically obtained audio chapters to the manually obtained audiovisual chapters in order to assess the effectiveness of the segmentation algorithm and to confirm our supposition that the audio stream conveys sufficient semantic information for a reliable segmentation.

Table 3 The precision (P %) and recall (R %) ratios of the automatically obtained chapters in comparison to manually obtained ones

	P %	R %	Mean boundary deviation error (s)	Chapters
Casino (1600 s)	77	70	3.5	24
Life is Beautiful (1600 s)	61	59	4.2	27
Training Day (1600 s)	70	64	4.3	22
Artificial Intelligence (1600 s)	53	47	8.5	19

A deviation of 8 s between true chapter boundaries and automatic ones was considered as acceptable in this experiment. In a video navigation application, this deviation is not problematic. The results are shown in Table 3. As we can see from the table, an average of 65 % of precision and 60 % of recall are obtained for the hard problem of chapter segmentation. The deviation between the automatically obtained chapter boundaries and the manually obtained ones is 5 s on average. Since, the basic attempts of the literature to segment a video into audio scenes include only some preliminary experimental results [7, 8, 10, 9], it is quite difficult to compare the proposed technique to those of the literature. Given the fuzzy definition of the chapters in movies due to the subjectivity engaged in a segmentation process, we can state that the obtained accuracies are acceptable and that the audio stream conveys indeed the semantic content of a video needed for the determination of the story units.

6.2 Classification

Speech/music/noise PGM–MLP classifier from the Sect. 3.2.1., was used for the classification of the automatically obtained chapters into the audio classes defined previously. A chapter is treated as a scene and the classification was done accordingly. The automatically obtained chapters, although not perfectly obtained were classified manually into

Table 4 The number correctly classified chapters and the number of chapters for each scene category (correctly classified chapters/total number of chapters)

	AD	CD	E	F	NA	SA	Total
Casino	1/1	6/7	1/3	0/0	0/0	2/4	10/15
Life is beautiful	8/9	2/3	0/1	0/0	2/2	0/0	12/15
Training day	3/4	6/7	1/1	1/2	1/1	0/0	13/15
A.I.	1/1	6/7	2/3	2/2	0/1	0/1	11/15

six classes, Active Dialog (AD), Calm Dialog (CD), Emotion (E), Fear (F), Natural Action (NA) and Special effects Action (SA). The Neutral dialog category was eliminated since it created confusion to the human subjects. In this experiment, Neutral dialog was fused with the Calm Dialog category. Table 4 shows the number of automatically given true labels and the manually given classes for the first 15 chapters of each movie corresponding to 1000 s of data per movie.

As we can see on the table, the mean classification accuracy is 78.7%. This accuracy seems acceptable for this hard problem of audio scene classification. However, it is important to notice that the classification results are based on relatively long audio segments, 60 s per segment. That is, the effect of local errors of the audio classifier is probably minimised when considering long audio segments. Nevertheless, the fact that the classification results are based on long audio segments, audio chapters, does not affect the usability of such classification from an application viewpoint.

A PGM–MLP audio classifier was also trained on the six audio classes directly and used on the same test data. That is the audio classifier was trained on the AD, CD, E, F, NA and SA audio classes. Surprisingly enough, the classification accuracy was considerably worse than that obtained using a basic Speech/music/noise classifier and rules. This is probably due to the non-suitability of the PGM features for modelling such audio classes.

7 Conclusion

We introduced in this paper the structuring of a video, a movie for instance, based on the audio-stream only. We showed how combining content information and acoustical information can provide a solution to the hard problem of semantic scene and chapter segmentation. A tree-like audio-based structure is obtained automatically where the content of each unit, a chapter for instance, is described. The proposed algorithm for chapter segmentation and the description of chapters into different scene categories were evaluated on four different movies. We can conclude that the audio analysis in movies can provide a solution for visualising the content. A user can hence have an on-the-fly idea about the content of a movie or a scene.

The combination with image analysis seems a natural future investigation. However, the information fusion needs further research especially that, as shown in this work, the audio structuring provides a different view on a video. That is, the currently used master(visual)–slave(audio) architecture needs to be studied.

Acknowledgements This work has been partially supported by the Cyrano project within the French RNRT program. The first author was partially supported by Grant number 3691-2001 from the French Ministry of Research and Technology. The authors would like to thank the anonymous reviewers whose critiques greatly improved the original manuscript.

Appendix

The references of the movies used in the evaluation section and the manually and automatically obtained scenes are presented in this appendix (Tables A.1 and A.2 and Figs. 10–13).

Table A.1 The references of the movies used in our experiments

Film	Année	Actors	Genre
Casino	1995	Robert De Niro, Joe Pesci	Drama/ crime
Training Day	2001	Denzel Washington, Ethan Hawke	Action
Life is Beautiful	1998	Roberto Benigni, Nicoletta Braschi,	Drama
Artificial Intelligence	2001	Haley Joel Osment, Jude Law	Science fiction

Table A.2 The chapter boundaries (in seconds) for the movies used in our experiments

Casino	Training Day	Artificial Intelligence	Life is beautiful
45	61	70	37
70	120	123	52
200	200	230	85
300	234	360	110
403	260	430	150
500	450	530	200
580	480	583	250
624	528	617	344
664	555	682	385
720	655	753	450
855	723	785	493
900	753	847	557
965	830	924	625
1035	850	1110	663
1078	940	1190	680
1140	1003	1237	810
1200	1080	1310	914
1260	1160	1450	965
1310	1250	1480	1005
1340	1295	1549	1060
1390	1460		1103
1430	1500		1180
1488			1245
1560			1290
			1380
			1465
			1522

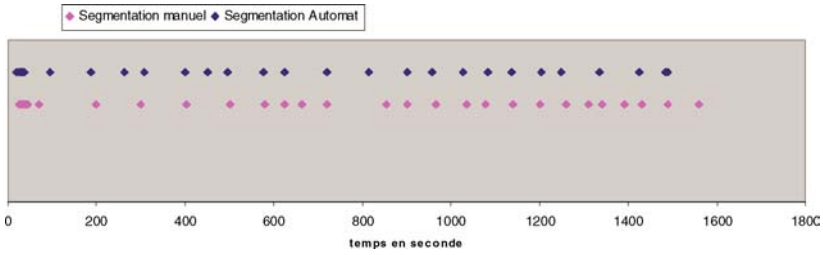

Fig. 10 Automatic and manual chapter boundaries "Casino"

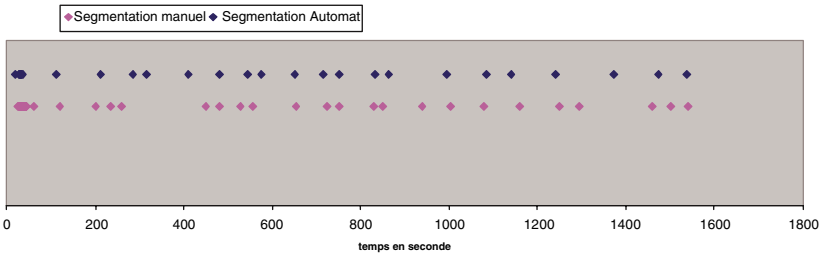

Fig. 11 Automatic and manual chapter boundaries "Training Day"

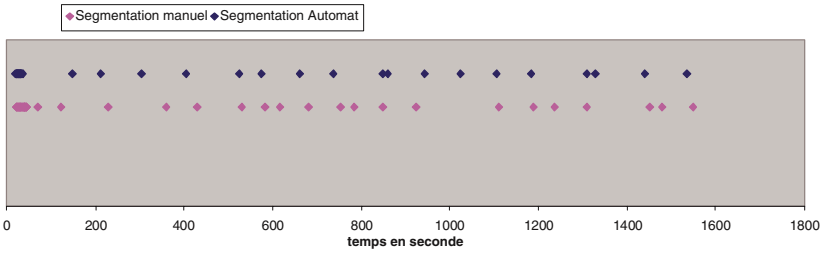

Fig. 12 Automatic and manual chapter boundaries "Artificial Intelligence"

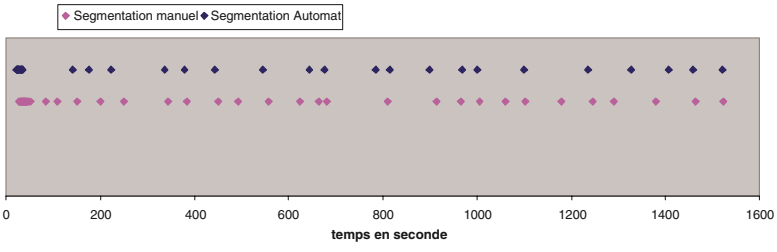

Fig. 13 Automatic and manual chapter boundaries "Life is Beautiful"

References

1. TREC video retrieval evaluation, [http://www-nlpir.nist.gov/projects/tv2004].
2. Gargi, U., Kasturi, R., Strayer, S.: Performance characterization of Video-Shot-Change detection methods. IEEE Trans. Circuits Syst. Video Technol. **10**(1):1–13 (2000)
3. Mahdi, W., Ardabilian, M., Chen, L.: Automatic video scene segmentation based on spatial-temporal clues and rhythm. Netw. Inform. Syst. J. **2**(5):1–25 (2000)
4. Fan, J., Elmagarmid, A., Zhu, X., Aref, W., Wu, L.: ClusterView: Hierarchical Video Shot Classification, indexing and accessing. IEEE Trans. Multimedia **6**(1):70–86 (2004)
5. Huang, C.L., Liao, B.Y.: A robust scene-change detection method for video segmentation. IEEE Trans. Circuits Syst. Video Technol. **11**(12):1281–1288 (2001)
6. Rui, Y., Huang, T., Mehrotra, S.: Constructing table-of-content for videos. Multimedia Syst. **7**(5):359–368 (1999)
7. Sundaram, H., Chang, S.F.: Audio scene segmentation using multiple features, models and time scales. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP, vol.4, pp. 2441–2444 (2000)
8. Cao, Y., Tavanapong, W., Kim, K., Oh, J.: Audio assisted scene segmentation for story browsing. Proceedings of International Conference on Imaging and Video Retrieval, Urbana-Chanpaign, IL, USA., pp. 446–455 (2003)
9. Minami, K., Akutsu, A., Hamada, H., Tomomura, Y.: Video handling with music and speech detection. IEEE Multimedia **5**(3):17–25 (1998)
10. Pfeiffer, S.: Scene determination based on video and audio features. Multimedia Tools Appl. **15**(1):59–81 (2001)
11. Chen, S.C., Shyu, M.L., Liao, W., Zhang, C.: Scene change detection by audio and video clues. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME2002), vol. 3, pp. 365–368 (2002)
12. Alatan, A., Akansu, A., Wolf, W.: Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing. Multimedia Tools Appl. **14**:137–151 (2001)
13. Saraceno, C., Leonardi, R.: Indexing audiovisual databases through joint audio and video processing. Int. J. Imaging Syst. Technol. **9**(5):320–331 (1998)
14. Harb, H., Chen, L.: Voice-based gender identification in multimedia applications. Journal of Intelligent Information Systems **24**(2–3): 179–198 (2005)
15. Gold, B., Morgan, N.: Speech and Audio Signal Processing: Processing and Perception of Speech and Music. Wiley, New York (1999)
16. Harb, H., Chen, L.: Highlights detection in sports videos based on audio analysis. Proceedings of the Third International Workshop on Content-Based Multimedia Indexing CBMI03, September 22–24, IRISA, Rennes, France, pp. 223–229 (2003)
17. Harb, H., Chen, L.: A Query by Example Music Retrieval Algorithm. Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services WIAMIS03, University of London, UK, 9–11 April, pp. 122–128 (2003)
18. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In: DARPA speech recognition workshop (1998)
19. Cover, T., Thomas, J.: Elements of Information Theory, Wiley Series in Telecommunications. Wiley, New York (1991)