On the Correlation of Automatic Audio and Visual Segmentations of Music Videos

Olivier Gillet, Slim Essid, and Gaël Richard, *Senior Member, IEEE*

***Abstract*—The study of the associations between audio and video content has numerous important applications in the fields of information retrieval and multimedia content authoring. In this work, we focus on music videos which exhibit a broad range of structural and semantic relationships between the music and the video content. To identify such relationships, a two-level automatic structuring of the music and the video is achieved separately. Note onsets are detected from the music signal, along with section changes. The latter is achieved by a novel algorithm which makes use of feature selection and statistical novelty detection approaches based on kernel methods. The video stream is independently segmented to detect changes in motion activity, as well as shot boundaries. Based on this two-level segmentation of both streams, four audio–visual correlation measures are computed. The usefulness of these correlation measures is illustrated by a query by video experiment on a 100 music video database, which also exhibits interesting genre dependencies.**

***Index Terms*—Audio segmentation, cross-modal queries, information retrieval, multimedia indexing, multimodal processing, music videos, novelty detection.**

I. Introduction

MULTIMEDIA document indexing refers to the process by which high-level descriptors or semantic representations are automatically extracted from documents. For example, such descriptors may take the form of a temporal structuring of the document in shots, a transcription of all spoken words or the detection of events of interest. Multimodal approaches combining audio, video, and possibly text streams have been successfully used for such tasks, for example to discover multimedia patterns and concepts [1], classify television programs [2] or identify interviewees in news broadcasts [3]. The multimodal dimension of multimedia documents is however not always straightforward to consider, and a large majority of studies focuses on unimodal indexing approaches.

Music-related audio–visual content (television broadcasts of concerts, operas or music videos) represents a specific class of multimedia data which is particularly interesting. In the case of music videos, a large palette of semantic relationships between the audio and video streams is used by the artists and directors. For instance, mainstream music videos show dancers or performers, some videos have a narrative content based on higher level features of the song (such as structure or mood); while others explore new forms of visual metaphors [4]–[6]. The interest for music video automatic processing is fairly recent and so far, mostly dedicated to automatic summarization (such summarization systems are described by Agnihotri *et al.* in [7] and [8] and Kankanhalli *et al.* in [9]).

In this work, which is an extension of a preliminary study [10], we further investigate the correlations of the audio and visual streams in music videos. For this purpose, four different correlation measures between the temporal structures of both streams are defined. Our main motivation is to describe how the video illustrates the music using these correlation measures. One of the main contributions of our work is to consider both streams at a structural level rather than at the feature level as it is traditionally done. An advantage of this approach is that it requires no prior media aesthetics knowledge regarding potential correlations between low level features of both streams. Such a matching of the audio and video content at a structural level opens the path for numerous applications, ranging from temporal resynchronization of mismatched audio and video streams to audio-driven video editing, or soundtrack retrieval by video query. Several systems have been developed to tackle such a retrieval problem. In [11], Foote *et al.* describe an audio-driven home-video summarization system, in which the highest quality segments of home videos are edited to match the structure of a target musical accompaniment. A video-driven system was developed by Dulhem *et al.* in [12]. Music pieces were retrieved to serve as a soundtrack to home videos by comparing the projections of global low-level audio and video features into a common pivot space. In [13], the correlation between tempo and motion activity is used to rank music pieces according to a video query. Finally, Nayak *et al.* describe in [14] a system generating MIDI background music constrained by low-level hue and brightness video features.

This paper is organized as follows. First, a brief overview of the structuring system is proposed in Section II. The next section is dedicated to audio event detection and audio segmentation. The video segmentation approaches selected are briefly described in Section IV. Section V introduces the audio–visual correlation measures derived from the automatic segmentations. Experimental results on a music video database are given in Section VI. Finally, Section VII suggests some conclusions and future directions.

II. Audio-Visual Content Analysis System

The aim of our system, whose overall architecture is given in Fig. 1, is to separately structure both audio and video streams,

Manuscript received July 11, 2006; revised November 9, 2006. This work was supported in part by the European Commission under the FP6-027026-K-SPACE Contract. This paper was recommended by Guest Editor E. Izquierdo.

The authors are with the LTCI-CNRS, GET-Télécom Paris, 75634 Paris, France (e-mail: olivier.gillet@enst.fr; slim.essid@enst.fr; gael.richard@enst.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2007.890831

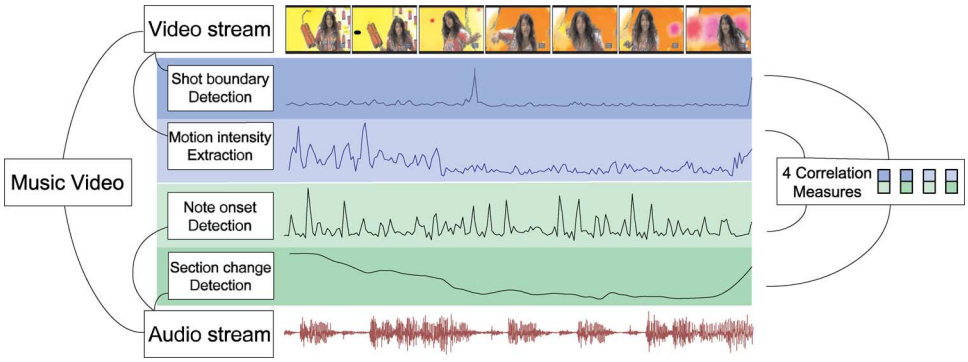

Fig. 1. Overview of the audio–visual content structuring system.

at two semantic levels, in order to measure the correlations between these structures. Hence, we try to characterize the synchrony of significant events and changes in the music and the accompanying images. This section defines in detail which events and changes are detected.

Some of the most salient events in music signals are note or chord changes. Thus, an efficient mid-level temporal structuring of a music piece can be achieved by detecting the onsets of such events which coarsely capture the rhythmic properties of the music. Since onset detection is a fundamental component of automatic music transcription and beat tracking systems, many approaches have been proposed to solve this problem (refer to [15] for a tutorial on the topic).

Likewise, the events of interest to be extracted from the video include rapid movements such as dance steps, movements of musicians or any action sequence. Such events can be efficiently detected by means of motion activity detectors [16].

At a higher level, a music piece can be temporally segmented in sections, characterized by distinct dynamic, tonal or timbral properties and corresponding to the musical structure of the piece, i.e., choruses, verses, fill-ins, etc. This problem is traditionally solved by computing a self-similarity matrix of the signal, and identifying large blocks within it, or by detecting boundaries between adjacent signal frames (such approaches are illustrated in [17], [18]). A very interesting alternative approach consists in using novelty detection methods which allow for determining boundaries between homogenous temporal segments. In this paper, this approach is further developed and a number of recent novelty detection methods are evaluated on this task.

At a higher level, the video stream is segmented into shots. In fact, shot changes events are semantically important in the sense that they may be correlated with the rhythm or section changes in the music. Shot detection is a topic that is widely studied. A review of state-of-the-art methods which had been evaluated in the latest TRECVID campaign, can be found in [19].

These four segmentation processes produce detection functions (represented in Fig. 1) ideally exhibiting peaks whenever an event or section change is detected. The detection functions can be thresholded to obtain the temporal location of salient events and segment boundaries, or directly considered to measure correlations.

The following sections present the algorithms used to perform the aforementioned mid-level and high-level structuring, for both the audio (Section III) and the video (Section IV) content.

III. Audio Processing

A. Audio Event Detection

Earlier methods for audio event detection in musical signals were solely based on the amplitude envelope of the waveform. It is now widely accepted that a more robust detection is achieved when the signal is first split into several frequency channels. However, there is no consensus on an optimal frequency decomposition for onset detection even if in most of the earlier studies a rather limited number of bands is used—for instance, the systems described in [20] and [21] use, respectively, 6 and 21 bands.

The onset detector implemented in this work is described in [22]. In order to accurately detect changes in the frequency content of the audio signal, a large number of frequency bands is used, in combination with the spectral energy flux (SEF) introduced in [23]. Firstly, the input signal $x(n)$ (where n is discrete time) is decomposed into 512 frequency channels using short-term Fourier transform (STFT), resulting in the so-called spectrogram $\hat{X}(m,k)$, m being a time frame index, and $k \in [0 \ldots 511]$ the frequency bin index. For each frequency bin k, the real, positive signal $|\hat{X}(m,k)|$ is low-pass filtered and its dynamic range is compressed, resulting in a perceptually plausible power envelope. Then, its derivative is computed by applying an optimal finite impulse response (FIR) differentiation filter, resulting in the spectral energy flux.

Finally, a detection function $d_o(m)$ is obtained by summing the SEF from each channel. This detection function typically exhibits sharp peaks at note onsets, chord changes or percussive events, and can be thresholded to obtain note onset times. In our case, we directly use $d_o(m)$ to investigate the correlations between audio and visual segmentations (see Section V).

B. Audio Section Change Detection

Audio section changes are discovered using a novelty detection module, based on statistical approaches, which use an efficient selection of audio features. We start by explaining how

these features are obtained, then briefly describe the various novelty detection techniques which we examined as well as the common framework in which they are used.

1) Feature Extraction: 70 candidate features are considered for this task. They are extracted over rather long windows of a 2-s length. Using long windows allows us to compensate for periodic and rhythmic variations of the features, hence model long-term phenomena. High overlap is kept between successive windows, as 8 frames are computed per second, in order to gather a large amount of data and increase the temporal accuracy of the decisions. The resulting representation is noted $X_f(m)$, where m is the frame index and f the feature index.

The features which were examined are briefly described hereafter. Interested readers are referred to [24] for a more detailed description.

a) Spectral Features: The spectral features include the following.

- Pitch class features [25]. These features describe the distribution of energy in 12 frequency bins corresponding to the 12 pitch classes of the equal-tempered scale.
- A subset of features is obtained from the statistical moments, namely the spectral centroid (from the first order moment), the spectral width (from the second order moment), the spectral asymmetry defined from the spectral skewness, and the spectral kurtosis describing the peakedness of the spectrum.
- Octave band signal intensities (OBSI) are computed to capture in a rough manner the power distribution of the different harmonics of a musical sound without resorting to pitch-detection techniques. Using a filterbank of overlapping octave band filters, the log energy of each subband (OBSI) and the logarithm of the energy ratio of each subband sb to the previous subband $sb-1$ (OBSIR) are measured [26].

b) Cepstral Features: Mel-frequency cepstral coefficients (MFCC) are extracted [27] to serve as an estimate of the spectral envelope of the signal.

c) Temporal Features:

- Zero crossing rates (ZCR) are extracted to help discriminating periodic signals (small ZCR values) from noisy signals (high ZCR values).
- The variance, skewness and kurtosis of the waveform over each observation window are measured. The same moments, along with the average, are also computed from the amplitude envelope of the waveform. To obtain the amplitude envelope, we first compute the modulus of the complex envelope of the signal, then filter it with a low-pass filter (which is the decreasing half of a 20 ms Hanning window). Such amplitude envelope features aim at modeling the rhythmic content.

d) Perceptual Features: Three perceptual features are extracted: the relative specific loudness (Ld) which corresponds to the loudness in each Bark band, the sharpness (Sh)—as a perceptual alternative to the spectral centroid based on specific loudness measures—and spread (Sp), which is the distance between the largest specific loudness and the total loudness [24].

2) Feature Selection: Feature selection arises from data mining problems where a subset of d features are to be selected from a larger set of D candidates, the selected subset being required to include the *most efficient* features. This issue has been extensively addressed in the statistical machine learning community [28], [29] and used for various classification tasks.

Nonsupervised feature selection techniques such as [30] simply aim at reducing the redundancy between features—*efficient* features are in this case the minimal, nonredundant subset that entirely describes the data. In supervised classification problems, when the classes are initially well defined, the features considered as *efficient* are generally the ones yielding the best classification performance. In these problems the selection criterion is hence related to the ability of the features to discriminate the considered classes. While our novelty detection problem is not a classification problem *per se*, this discriminative approach is still valid: we aim at extracting a set of features that will discriminate frames from two distinct sections, but not frames drawn from the same section. In other words, we can consider that pairs of adjacent audio sections are two distinct classes to discriminate.

Thus, we use the following semi-supervised approach.

- First, a hold-out set of music signals is manually segmented according to their structure.
- Then, each segment boundary in every signal is considered to define a biclass classification problem—considering the "past data class" and the "future data class" related to each boundary.
- For each biclass problem Π_i defined as above, we apply a supervised feature selection procedure. We chose to use a simple approach, based on Fisher's linear discriminant algorithm (LDA) [31], which computes the relevance of each candidate feature using the weights estimated by the LDA. In fact, extensive experimentation had been carried out in previous work [32] comparing various feature selection techniques, which motivated this choice. Both "filter" algorithms (which use the initial set of features intrinsically), and "wrapper" algorithms (which relate the selection procedure to the performance of the classifiers considered) had been envisaged. Hence, we worked out that very simple filter algorithms, particularly LDA selection, always produce appropriate subsets of features with the advantage of being computationally inexpensive.
- The previous step yields several possible subsets of selected features (one subset per biclass problem Π_i). We then rely on a voting procedure to produce a unique subset of selected features to be used for segmenting all signals: each time a given feature is found in the subset of selected features for problem Π_i, it receives a vote, then features are ranked again with respect to the number of votes they received.

Thirty-two features have thus been kept from the original set of 70 candidates. This target number of features has been selected after testing on the hold-out set. For validation purposes, we divided our music signal database in two parts (each amounting to 50% of the total size of the database) and performed the feature selection procedure described above on each part—one part being considered as a hold-out set for the other in the evaluation stage. The selected features in both hold-out sets were the same, which suggests that the best feature set ob-

TABLE I
SELECTED FEATURES

Features group	Selected features	Total number of features
Pitch class	0	12
Spectral moments	2	4
OBSI	7	7
MFCC	3	13
ZCR	1	1
Waveform moments	2	3
Envelope moments	2	4
Perceptual	13	26

tained (and given in Table I) is quite stable. We, therefore, used it in the evaluation of novelty detection algorithms, as described hereafter.

3) Novelty Detection Techniques: The novelty detection problem can be stated as follows: given a set of reference examples, decide whether a set of observations is generated by the same process as the one underlying the reference examples. The segmentation of observation sequences is an instance of this problem. In fact, deciding whether or not a section change occurs at time t_0 is akin to deciding whether the observations at time $t > t_0$ (*future data set*) are novel with respect to the observations at time $t < t_0$ (*past data set*). In practice, only a limited number of observations are considered for the past and future data sets. All the novelty detection approaches used in this work are thus based on this same problem formulation: a sliding window $W(m_0)$ of length $2L+1$ centered at frame m_0 is observed. m_0 is considered as a good candidate for being a segment boundary if the content of the future data set $S_2(m_0) = \{X(m), m \in [m_0, m_0 + L]\}$ is novel relatively to the content of the past data set $S_1(m_0) = \{X(m), m \in [m_0 - L, m_0]\}$, where $X(m)$ is the feature vector measured on frame m. To simplify notations, the past and future windows, for a given value of m_0 will be simply referred to as S_1 and S_2, the underlying probability distributions as P_1 and P_2, the entire window as W, the feature vector as X.

Solutions to the novelty detection problem typically differ in the class of models used for estimating the distributions P_1 and P_2; and in the criterion used to compare them. The three methods given here illustrate this variety of solutions.

a) Bayesian Information Criterion (BIC): Being a classical model or order selection criterion, the BIC has been widely used in speech/music or speakers segmentation problems [33], [34]. Hence, it will be considered here as a baseline algorithm.

We assume the elements of S_i to be distributed according to P_i if the considered value of m_0 is a segment boundary, otherwise (if m_0 is not a segment boundary), the elements of the entire observation window W are assumed to be distributed according to a single distribution P.

In the case of Gaussian distributions, the BIC variation between the two models can be expressed as

$$\Delta\mathrm{BIC} = \frac{1}{2}\left((2L+1)\log|\Sigma| - L\left(\log|\Sigma_1| - \log|\Sigma_2|\right) - \kappa\right)$$

where the covariance matrices Σ_i and Σ are, respectively, estimated from S_i and W. Since we are only interested in finding local maxima of $\Delta\mathrm{BIC}$, the constant κ does not need to be explicited here.

b) Novelty Detection With One-Class Support Vector Machines: One-class support vector machines (SVM) aim at identifying a region of the feature space in which most of the data points lie. This is obtained by finding the hyperplane that separates the data from the origin with maximum margin [35].

Two different novelty detection approaches based on one-class SVM are considered in this work: the first is based on a likelihood ratio test [36] and the second is the so-called kernel change detection (KCD) approach [37]. We will only explain how the novelty detection criteria are computed in each case. For further details on these techniques, we refer the reader to [35]–[37].

A likelihood ratio test can be defined as

$$R = \frac{\prod_{X\in S_1} P_1(X)\prod_{X\in S_2} P_2(X)}{\prod_{X\in W} P_1(X)} = \frac{\prod_{X\in S_2} P_2(X)}{\prod_{X\in S_2} P_1(X)} > t.$$

Hence, estimates of P_1 and P_2 are needed in order to perform this test, which can be easily deduced from the SVM algorithm solution $\{\eta_i, \alpha_m^i\}$, according to

$$\hat{P}_i(X) = \exp\left(\sum_m \alpha_m^i K\left(X, X^i(m)\right) - \eta_i\right)$$

where η_i is a threshold, α_m^i are Lagrange multipliers, $(X^i(m))_m = S_i$ are the vectors of the training set, and K is a reproducing kernel.

The numerator of R indicates how well the 1-class SVM algorithm fits its own training set, and is expected to be close to one. Hence, the detection criterion can be simplified as

$$R' = \frac{1}{\prod_{X\in S_2} \hat{P}_1(X)} > t.$$

Let $\mathbf{K}_{ij}$ be the kernel matrix with elements $K(X^i(m), X^j(l))$ at the mth row and lth column, with $(i,j) \in \{1,2\}\times\{1,2\}$ and $X^i(m)$ the mth training vector of the set S_i. The KCD approach is based on a dissimilarity measure that can be seen as a ratio of *inter-class scatter* to *intra-class scatter* in the transformed space induced by the kernel. This dissimilarity is defined as follows:

$$\mathcal{D} = \frac{\widehat{c_1c_2}}{\widehat{c_1p_1} + \widehat{c_2p_2}}$$

where

$$\widehat{c_1c_2} = \arccos\left(\frac{\boldsymbol{\alpha}_1^T\mathbf{K}_{12}\boldsymbol{\alpha}_2}{\sqrt{\boldsymbol{\alpha}_1^T\mathbf{K}_{11}\boldsymbol{\alpha}_1}\sqrt{\boldsymbol{\alpha}_2^T\mathbf{K}_{22}\boldsymbol{\alpha}_2}}\right)$$

and

$$\widehat{c_ip_i} = \arccos\left(\frac{\eta_i}{\sqrt{\boldsymbol{\alpha}_i^T\mathbf{K}_{ii}\boldsymbol{\alpha}_i}}\right), \quad i \in \{1,2\}.$$

It is worth mentioning that it is not necessary, for both methods, to run the 1-class SVM algorithm entirely over each new observation window. In fact, since $S_1(m_0)$ and $S_1(m_0+1)$ share L data points in common, one can merely remove the

outgoing points from the set of support vectors (if necessary), and perform the optimization starting with the existing set of support vectors. Furthermore, both expressions $\mathcal{D}$ and R only depend on the set of support vectors and the corresponding Lagrange multipliers. These two observations allow us to achieve a substantial reduction of the computational load.

c) Probabilistic Distances: Another way of detecting segment boundaries is by using a relevant distance between the data points in S_1 and S_2. We expect these boundaries to be characterized by a higher distance. For the sake of robustness we consider probabilistic distances between the estimates of the distributions $\hat{P}_1$ and $\hat{P}_2$. Many such distances can be considered among which we chose the Bhattacharya distance and the Kullback-Leibler divergence (mainly due to the resulting simplification in the following computations).

While these distances admit analytical expressions whenever the probability densities are Gaussian, computing them can be otherwise a difficult problem since it requires performing heavy numerical integrations [38]. In fact, in the Gaussian case, the distances can be expressed as functions of the means and covariance matrices of the multivariate Gaussian densities describing, respectively, class 1 and class 2 in $\mathbb{R}^D$. Nevertheless, it would be highly suboptimal, in our case, to assume that the original class observations follow Gaussian distributions since we deal with data with a nonlinear structure. Fortunately, if this data is mapped from the original space to a reproducing kernel Hilbert space (RKHS) [35], it is reasonable to assume it to be Gaussian in the RKHS [38].[1] Thus, a robust estimation of the needed probabilistic distances can be derived using analytical expressions provided that a proper estimation of the means and covariance matrices in the RKHS can be obtained. The strength of such an approach resides in that there is no need for explicitly knowing either the structure of the original probability densities or the nonlinear mapping to be used. Interested readers are referred to [38] for further details.

4) Normalization and Thresholding: Because of their large dynamic range, the novelty detection functions $d(m)$, output of the BIC, 1-class SVM or probabilistic distance algorithms need further processing to ease section change detection. Two nonlinear filters are applied to them: firstly, the detection function $d(m)$ is detrended by removing a median-filtered version of itself. A window size of M_l is used for this median filter. Then, local variations in peaks amplitude are compensated by dividing the resulting detrended detection function $d_d(m)$ by a standard-deviation filtered version of itself

$$d_c(m)=\frac{d_d(m)}{\operatorname{std}\left[d_d\left(m-\frac{M_l-1}{2}\right),\ldots,d_d(m),\ldots,d_d\left(m+\frac{M_l-1}{2}\right)\right]}$$

Local maxima above a given threshold τ in the function $d_c(m)$ indicate section changes. Such local maxima can be easily identified by defining the following "top-hat" function $d_t(m)=\max[d_c(m-((M_s-1)/2)),\ldots,d_c(m),\ldots,d_c(m+((M_s-1)/2)),\tau]$ and by detecting a section boundary whenever $d_t(m)=d_c(m)$. Practically, this detection process is such that the length of the detected sections lies within the range $[M_s,M_l]$. Therefore, we used in this work $M_s=40$ frames (5 s); and $M_l=360$ frames (45 s), which correspond to minimum and maximum section lengths observed in our database.

[1]This assumption is the basis of methods such as kernel-PCA or kernel-Fisher discriminant analysis.

IV. Video Processing

A. Detection of Video Event Onsets

The mid-level structuring of the video stream consists in identifying the onset times of significant changes in motion. Such changes can correspond, in the case of music videos, to musicians movements to play notes, to dance moves, or to action sequences. While video tracking systems are available for specific music related activities such as drumming [39] or dancing [40], these systems only work in well-controlled environments, with fixed cameras—which make them unsuitable for overall motion analysis in generic music videos. Thus, lower-level general-purpose motion features are used, based on the MPEG-7 motion activity descriptor [16].

Motion vectors are available from the P- (predicted) frames of MPEG video streams. Motion vectors associated with nontextured blocks, on which motion estimation is inaccurate, are discarded. The motion vector field is smoothed by a 3×3 median filter. Then, we extract a motion activity feature, corresponding to the standard deviation of the motion vectors' magnitude. The use of standard deviation increases the robustness of this feature to continuous motion components induced by camera motions (such as panning). Finally, changes in this motion feature are detected by using a differentiating filter, resulting in a detection function $d_m(m)$.

B. Detection of Shot Boundaries

At a higher level, the video stream is segmented in shots. State-of-the-art algorithms for this task are reviewed in the latest (2005) annual TRECVid evaluation report [19].

In the case of music videos, the shot boundary detection is greatly simplified by the fact that transitions between shots are mostly cuts: we have observed in a 30 music videos subset of our corpus that 91% of the transitions are cuts. Moreover, as we ultimately aim at correlating the video and audio streams, false positives in cut detection such as flashing lights and fast lighting changes are tolerable, as these effects are often beat-synchronous, and are thus worth being detected. By contrast, dissolves and fades are less localized in time, and thus harder to correlate with audio events.

Consequently, we use a simple cut detection system based on the distance of color and luminosity features between adjacent frames. Three 16-bins histograms are computed for each frame, from the Y, U and V components. This results in a 48 features vector $X(m)$ for each frame. The shot boundary detection function is defined as

$$d_s(m)=\|X(m)-X(m-1)\|_1=\sum_{q=1}^{48}|X_q(m)-X_q(m-1)|$$

V. Audio–Visual Correlation Measures

A. Overview

A possible procedure to correlate the structure of the video and audio streams could be to match the corresponding segmentations—for example by computing an edit distance between them (possibly by counting the number of split or merge operations), or by counting the number of changes occurring simultaneously in both streams. However, we did not follow this approach for several reasons. Firstly, it requires the definition of a decision threshold for each detection function, which can suppress meaningful section changes. Secondly, it does not take into account the salience of each event.

For these two reasons, we decided to directly correlate the detection functions, rather than the structures extracted from them. Thus, we consider the note onset detection function $d_o(m)$, the detrended section change detection function $d_c(m)$, the shot boundary detection function $d_s(m)$, and the motion activity changes function $d_m(m)$. All these detection functions are centered, normalized by their standard deviations and converted to a common sampling rate (25 Hz, which corresponds to the frame rate of the videos used in the experiments).

B. Time Warping

We aim at characterizing whether changes in the audio stream and video stream—be it at a mid- or high-level—are perceived as simultaneous. This problem is thus equivalent to verifying whether peaks in the audio and video change detection functions occur simultaneously. However, changes perceived as simultaneous are not necessarily perfectly synchronized, due to the filters used in the computation of the detection functions; or imprecisions in the editing. Thus, prior to the computation of the correlation measures, the audio and video detection functions are aligned by means of the classical dynamic time warping (DTW) algorithm [41] to maximize the simultaneity of their peaks. The DTW is constrained to only search for alignment paths in the neighborhood of the diagonal (upper and lower diagonals). That is to say, only audio and video events differing by less than 2 frames (80 ms) are matched and thus considered as simultaneous.

C. Correlation Measures

Various statistical or information-theoretic measures can be used to correlate two detection functions $a(m)$ and $b(m)$. We can assume that $a(m)$ and $b(m)$ are sequences of independent realizations of a random variable A and B, and consider:

- Pearson's correlation, defined as

$$\rho(A,B) = \frac{\mathbb{E}\left[(A-\mathbb{E}[A])\,(B-\mathbb{E}[B])\right]}{\sqrt{\mathbb{E}\left[(A-\mathbb{E}[A])^2\right]\mathbb{E}\left[(B-\mathbb{E}[B])^2\right]}}$$

  where $\mathbb{E}$ is the statistical expectation.
- Mutual information, defined in the discrete case as

$$I(A,B) = \sum_a \sum_b P(A,B)\log\frac{P(A,B)}{P(A)P(B)}$$

  To allow the computation of this quantity, the values of $a(m)$ and $b(m)$ are discretized into 32 bins, using a standard histogram procedure.

Since both streams are segmented at two levels (onsets and sections for audio, motion and shots for video), four correlations can be defined. For each of them, out of the two possible measures—Pearson's correlation and mutual information—we selected the one that maximized the performance of our system for the retrieval experiment in Section VI-B.

Thus, the four audio–visual correlations measures used are the following:

$$C_{\text{onsets/shots}} = \rho(d_o, d_s) \quad (1)$$

$$C_{\text{sections/shots}} = \rho(d_c, d_s) \quad (2)$$

$$C_{\text{onsets/motion}} = I(d_o, d_m) \quad (3)$$

$$C_{\text{sections/motion}} = \rho(d_c, d_m). \quad (4)$$

VI. Experimental Results

A. Evaluation of the Audio Section Segmentation

As the audio segmentation system introduced in Section III-B is novel, it is subjected in this section to an independent evaluation. A database of 100 full-length pop music signals, (60 of them having been used in our previous work [10]) has been manually segmented. This database is subsequently referred to as Music-100.

Detection functions are computed for all the signals, with the features set given in Section III-B1. The 100 detection functions are thresholded as described in Section III-B4 with 70 different values of τ ranging from -2 to 5. Detected segment boundaries are classified as correct if they occur within a 4-s time window centered at each ground-truth segment boundary. Standard precision and recall scores are thus computed for each value of τ, yielding the curves in Fig. 2

$$\begin{aligned}\text{Precision} &= \frac{\text{Number of correctly detected boundaries}}{\text{Total number of detected boundaries}}\\ \text{Recall} &= \frac{\text{Number of correctly detected boundaries}}{\text{Number of boundaries to be detected}}\\ \text{F-measure} &= \frac{2\cdot\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}.\end{aligned}$$

Additionally, F-measure scores are computed with this last expression for a fixed value of $\tau = 1$, and given in Table II. The best results are obtained with the method based on Bhattacharya distance in RKHS. The kernel change detection algorithm and KL-divergence methods are less accurate for a given range of recall rates. The results obtained with the BIC are significantly worse. This can be explained by the fact that this criterion relies on the hypothesis that the data is Gaussian, which is not valid in our case. Using mixtures of Gaussians (GMM) to model the future and past distributions could have overcome this limitation, however, in our case, the lack of data from the observation windows would not have allowed us to robustly train such models. A more tractable approach adapted to the small observation windows used in our work could consist in using GMM adaptation of generic models—defined for example for each music genre, or instrumentation. However, this would have reduced

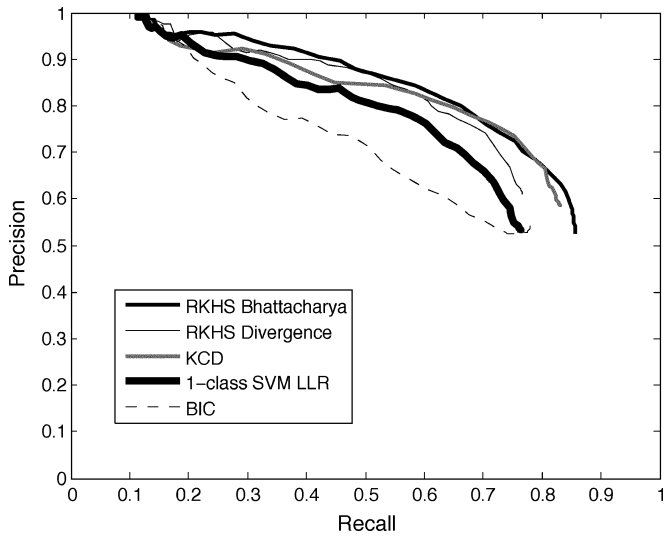

Fig. 2. Recall-Precision curves for the audio segmentation task on the Music-100 database: comparison of the different novelty detection algorithms.

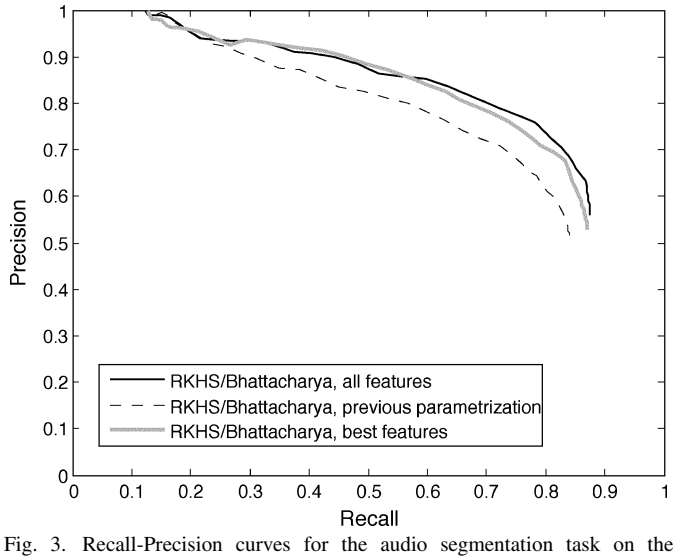

Fig. 3. Recall-Precision curves for the audio segmentation task on the Music-100 database: comparison of different features set.

TABLE II
F-MEASURE SCORES FOR DIFFERENT NOVELTY DETECTION ALGORITHMS, ON THE MUSIC-100 DATABASE

Algorithm	**F-measure for $\tau = 1$**
Bhattacharya distance in RKHS	0.74
KL Divergence in RKHS	0.68
Kernel Change Detection	0.72
1-class SVM, Likelihood ratio test	0.67
Bayesian Information Criterion	0.59

the ability of our algorithm to deal with unusual timbral textures or music genres. The failure of such generative approaches highlights the robustness and relevance of kernel methods for high-dimensional problems with scarce observations.

Precision and recall curves are also given in Fig. 3 for the best method (Bhattacharya distance in RKHS), with different sets of features: the whole features set computed in Section III-B1, the 32 features obtained with the feature selection process, and the previous parametrization described in [10]. The new features set leads to better performance compared to our previous parametrization. Moreover, there is no significant difference of performance between the entire features set and the reduced set obtained after feature selection. Feature selection can thus be seen as a way of reducing the complexity of the novelty detection process without causing performance degradations. It is also important to mention that the 70 candidate features introduced in III-B1 were preselected by the authors for this segmentation application. An additional experiment employing a larger and more diverse set of candidates may provide stronger arguments to justify the importance of feature selection.

B. Audio Retrieval From Video

In this second experiment, we consider the problem of finding the music corresponding to a given video, using the correlation measures defined above. For this purpose, we gathered a database of 100 music videos (Videos-100), from various sources: 25 music videos of high aesthetic and production value from [4]–[6], and 75 music videos aired on the MTV and MCM television channels. All the videos are encoded in the MPEG-2 format, with a resolution of 320 × 240, at 25 fps.

Given a video sequence V_i, $i \in [1 \dots 100]$, taken from the database, all the audio streams $(A_j)_{j\in[1...100]}$ available in the database are ranked according to one of the audio–visual correlation measures defined above. A threshold θ is used to define the set $R_i(\theta)$ of audio tracks which are the most correlated with the video query V_i

$$R_i(\theta) = \{j, C(A_j, V_i) > \theta\}\,.$$

For this retrieval experiment using a video query, precision and recall are defined as

$$\text{Precision}_i(\theta) = \begin{cases} \frac{1}{|R_i(\theta)|}, & \text{if } i \in R_i \\ 0, & \text{if } i \notin R_i \end{cases}$$

$$\text{Recall}_i(\theta) = \begin{cases} 1, & \text{if } i \in R_i \\ 0, & \text{if } i \notin R_i \end{cases}$$

where $|\cdot|$ denotes set cardinality. Global Precision and Recall scores, for a given value of θ, are obtained by averaging $\text{Precision}_i(\theta)$ (resp. $\text{Recall}_i(\theta)$), $i \in [1 \dots 100]$. Recall-Precision curves are given in Fig. 4.

The fast decrease of precision, as recall increases, suggests that the correlations are relevant for retrieval only on a subset of the database. Among this subset, the best performance can be achieved by considering the correlation between shot changes and note onsets. Yang and Brown performed in [13] a related retrieval experiment, by correlating a beat-tracking detection function, and motion features. They obtained a perfect match for each of their five test video queries. However, the nature of the material used in their work and the size of their dataset does not allow us to make a direct comparison.

It is worth noting that according to this evaluation protocol, for a given video sequence, the only piece of music counted as a correct result is the piece of music for which the video was produced. Highly correlated pairs of video and audio segments extracted from two distinct music videos are counted as errors.

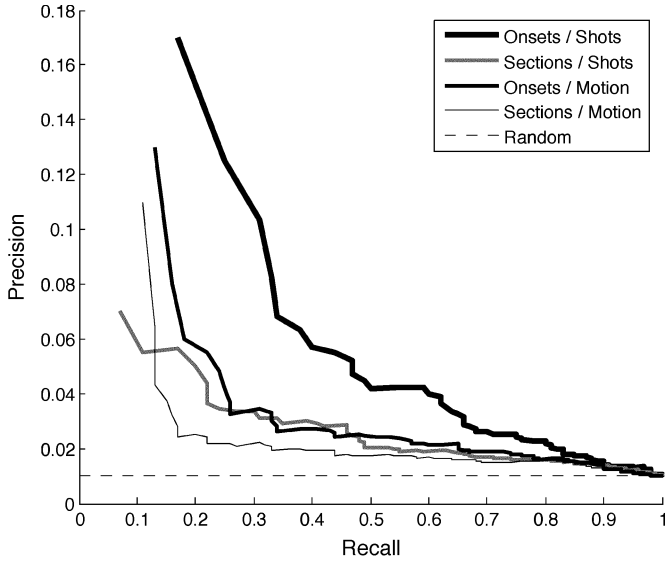

Fig. 4. Recall-Precision curves for the music retrieval with video query experiment, on the Videos-100 database.

Nevertheless, some of these pairs can be meaningful and sometimes surprisingly interesting—suggesting that some results considered as incorrect could indeed be relevant to a human user.

C. Influence of Music Video Genre on Audio–Visual Correlations

Prior to the following experiment, music videos were manually classified into five categories, depending on their video content and its relationship to the music.

- **Narrative**: The music video has a strong narrative content and chronology.
- **Musicians**: The music video mostly features shots of the band members performing the music.
- **Dance**: The music video mostly shows dancers (this can include the singer) dancing to the music.
- **Abstract visuals**: The music video is a sequence of shots with a high-level semantic relationship with the music. For example shots are related to the song lyrics, mood or atmosphere.
- **Video sampling, VJing**: The music video is made of short video clips edited or triggered to match the beat and samples used in the music.

When several categories could be used to label a music video, the category describing the largest number of sequences was used.

The aim of the experiment is to identify for which categories the correlation measures previously defined are relevant and useful for retrieval. For this purpose, for each video stream V_i, all the audio streams A_i are ranked according to their correlation with V_i. Let r_i be the rank assigned to the original audio stream that accompanied the video V_i. Low values of r_i indicate that the relationship between the music and visuals is strong enough to allow the music to be easily retrieved from the video content. Table III lists the average of r_i in each of the five subsets presented above.

It can be seen that the audio–visual correlations are extremely significant for the video sampling category. More generally, these correlations are more efficient for the retrieval of music videos emphasizing on music related activities (dance, performances) than on *narrative*, or *abstract visuals* videos for which the audio and video can only be matched efficiently at a higher semantic level.

TABLE III
INFLUENCE OF MUSIC VIDEO GENRE ON RETRIEVAL RESULTS

Category	Average rank of the original video
Narrative	23
Abstract visuals	19
Dance	13
Musicians	11
Video sampling	6

VII. CONCLUSION AND FUTURE WORK

Segmentation and event detection algorithms for music and video content processing were presented. In particular, we introduced a novel approach for music segmentation that makes use of feature selection and statistical novelty detection algorithms based on kernel methods. This novel approach was individually evaluated on a database of pop music signals, and showed improvements over the baseline BIC algorithm.

Audio–visual correlation measures were derived from the segmentations, allowing for the detection of co-occuring changes in the audio and video content of music videos. These correlations can be used to match audio and video content for retrieval applications, especially audio retrieval from video query. This was validated by an experimental evaluation on a corpus of 100 music videos and gave promising results. The results show that the performance depend on the music video genre, hence suggesting genre classification applications.

Future work will address this genre classification problem, by defining additional video and audio features, as well as new forms of correlations. The matching of the audio and video streams could also be enhanced by considering intermediate or higher levels of segmentation—for example by detecting beats (rather than note onsets) in the music, or by identifying sequence changes in the video. Further applications and developments of our work could also include automatic generation of music videos, or query by video systems to assist audio mixing and soundtrack composition.

REFERENCES

[1] L. Xie, L. Kennedy, S.-F. Chang, A. Divakaran, H. Sun, and C.-Y. Lin, "Discovering meaningful multimedia patterns with audio–visual concepts and associated text," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2004, vol. 4, pp. 2383–2386.
[2] J. Huang, Z. Liu, and Y. Wang, "Classification of TV programs based on audio information using hidden Markov model," in *IEEE Workshop Multimedia Signal Process.*, 1998, pp. 27–32.
[3] A. Albiol, L. Torres, and E. Delp, "Combining audio and video for video sequence indexing applications," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2002, pp. 353–356.
[4] M. Gondry (director), *The Work of Director Michel Gondry.* DVD, Palm Pictures, 2003.
[5] S. Jonze (director), *The Work of Director Spike Jonze.* DVD, Palm Pictures, 2003.
[6] EAF Music *Visual Niches—Extraordinary Music Videos*, DVD, , 2002.
[7] L. Agnihotri, N. Dimitrova, J. Kender, and J. Zimmerman, "Music videos miner," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 442–443.

[8] L. Agnihotri, N. Dimitrova, and J. R. Kender, "Design and evaluation of a music video summarization system," in *Proc. 2004 IEEE Int. Conf. Multimedia Expo*, Jun. 2004, pp. 1943–1946.
[9] X. Shao, C. Xu, and M. S. Kankanhalli, "A new approach to automatic music video summarization," in *Proc. Int. Conf. Imag. Proc.*, Oct. 2004, pp. 625–628.
[10] O. Gillet and G. Richard, "Comparing audio and video segmentations for music videos indexing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006.
[11] J. Foote, M. Cooper, and A. Girgensohn, "Creating music videos using automatic media analysis," in *Proc. 10th ACM Int. Conf. Multimedia*, 2002, pp. 553–560.
[12] P. Mulhem, M. S. Kankanhalli, J. Yi, and H. Hassan, "Pivot vector space approach for audio–video mixing," *IEEE Multimedia*, vol. 10, no. 2, pp. 28–40, Apr./Jun. 2003.
[13] R. Yang and M. S. Brown, "Music database query with video by synesthesia observation," in *Proc. 2004 IEEE Int. Conf. Multimedia Expo*, Jun. 2004, pp. 305–308.
[14] M. Nayak, S. H. Srinivasan, and M. S. Kankanhalli, "Music synthesis for home videos: An analogy based approach," in *Proc. IEEE Pacific-Rim Conf. Multimedia*, Dec. 2003.
[15] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 9, pp. 1035–1047, Sep. 2005.
[16] S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 720–724, Nov. 2001.
[17] G. Peeters, A. L. Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proc. 3rd Int. Conf. Music Information Retrieval*, 2002, pp. 86–92.
[18] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proc. Int. Symp. Music Inf. Retrieval*, 2002, pp. 81–85.
[19] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton, "TRECVID 2005—An overview," National Institute of Standards and Technology (NIST), 2006 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tv5overview.pdf
[20] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, 1998.
[21] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 3089–3092.
[22] M. Alonso, G. Richard, and B. David, "Extracting note onsets from musical recordings," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, CD-ROM.
[23] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *J. Audio Eng. Soc.*, vol. 51, no. 4, pp. 226–233, Apr. 2004.
[24] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," IRCAM, 2004.
[25] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. 2001 IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 15–18.
[26] S. Essid, G. Richard, and B. David, "Musical instrument recognition based on class pairwise feature selection," in *Proc. 5th Int. Conf. Music Information Retrieval*, Oct. 2004, pp. 560–567.
[27] L. R. Rabiner, *Fundamentals of Speech Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[28] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.
[29] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1–2, pp. 245–271, 1997.
[30] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.
[31] R. Duda, P. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2001.
[32] S. Essid, "Classification automatique des signaux audiofrequences: Reconnaissance des instruments de musique," Ph.D. dissertation, Université Pierre et Marie Curie, Paris, France, 2005.
[33] B. Zhou and J. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion," in *Proc. Int. Conf. Spoken Language Process.*, 2000, pp. 714–717.
[34] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcrip. Understand. Workshop*, Feb. 1998, pp. 602–610.
[35] B. Shölkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.
[36] G. Loosli, S. G. Lee, and S. Canu, "Context changes detection by one-class SVMs," in *Proc. Workshop Mach. Learning for User Modeling*, 2005, pp. 27–34.
[37] F. Desobry, M. Davy, and C. Doncarli, "An online Kernel Change Detection algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2961–2974, Aug. 2005.
[38] S. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing Kernel Hilbert Space," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 917–929, Jun. 2006.
[39] O. Gillet and G. Richard, "Automatic transcription of drum sequences using audiovisual features," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Mar. 2005, pp. 205–208.
[40] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe, "Multimodal analysis of expressive gesture in music and dance performances," in *Proc. 5th Int. Gesture Workshop*, Apr. 2003, pp. 20–39.
[41] J. B. Kruskal, "An overview of sequence comparison," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff and J. B. Kruskal, Eds. Reading, MA: Addison-Wesley, 1983, pp. 1–44.

Olivier Gillet received the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, and the M.Sc. (DEA) degree in artificial intelligence and pattern recognition from the Université Pierre et Marie Curie Paris, France, both in 2003. He is currently working toward the Ph.D. degree at the Department of Signal and Image Processing (TSI), ENST.

His research interests include signal processing and machine learning for audio content analysis, and the integration of video information into music information retrieval systems.

Slim Essid received the state engineering degree from the École Nationale d'Ingénieurs de Tunis, Tunisia, in 2001, the M.Sc. (D.E.A.) degree in digital communication systems from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2002, and the Ph.D. degree from the Université Pierre et Marie Curie, Paris, France, in 2005, after completing a thesis on automatic audio classification with the Department of Signal and Image Processing (TSI), ENST.

Currently, he is a Teacher and Research Engineer at TSI, ENST. His research interests include signal processing and machine learning for multimedia indexing, especially joint audiovisual processing, and music information retrieval (MIR).

Gaël Richard (M'02–SM'06) received the State Engineering degree from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1990, the Ph.D. degree in speech synthesis from LIMSI-CNRS, University of Paris-XI, Paris, France, in 1994, and the Habilitation à Diriger des Recherches degree from the University of Paris XI, Paris, France, in September 2001.

He spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the speech processing group of Prof. J. Flanagan, where he explored innovative approaches for speech production. Between 1997 and 2001, he successively worked for Matra Nortel Communications, Bois d'Arcy, France, and for Philips Consumer Comunications, Montrouge, France. In particular, he was the Project Manager of several large-scale European projects in the field of multimodal verification and speech processing. In September 2001, he joined the Department of Signal and Image Processing, GET-Télécom Paris (ENST), where he is now Full Professor in the field of audio and multimedia signals processing. He is coauthor of over 50 papers and inventor in a number of patents, he is also one of the expert of the European commission in the field of man/machine interfaces.