Sound Indexing Using Morphological Description

Geoffroy Peeters and Emmanuel Deruty

***Abstract*—Sound sample indexing usually deals with the recognition of the source/cause that has produced the sound. For abstract sounds, sound-effects, unnatural or synthetic sounds this cause is usually unknown or unrecognizable. An efficient description of these sounds has been proposed by Schaeffer under the name morphological description. Part of this description consists in describing a sound by identifying the temporal evolution of its acoustic properties to a set of profiles. In this work, we consider three morphological descriptions: dynamic profiles (ascending, descending, ascending/descending, stable, impulsive), melodic profiles (up, down, stable, up/down, down/up) and complex-iterative sound description (non-iterative, iterative, grain, repetition). We study the automatic indexing of a sound into these profiles. Because this automatic indexing is difficult using standard audio features, we propose new audio features to perform this task. The dynamic profiles are estimated by modeling the loudness over-time of a sound by a second-order B-spline model and derive features from this model. The melodic profiles are estimated by tracking over time the perceptual filter which has the maximum excitation. A function is derived from this track which is then modeled using a second-order B-spline model. The features are again derived from the B-spline model. The description of complex-iterative sounds is obtained by estimating the amount of repetition and the period of the repetition. These are obtained by computing an audio similarity function derived from an MFCC similarity matrix. The proposed audio features are then tested for automatic classification. We consider three classification tasks corresponding to the three profiles. In each case, the results are compared with the ones obtained using standard audio features.**

***Index Terms*—Sound description, automatic indexing, audio features, loudness, audio similarity.**

I. Introduction

MOST of the research in sound description focuses on the recognition of the sound source (the cause that has produced the recorded sound). For example [1] [2] [3], [4] propose systems for the automatic recognition of musical instruments (the cause of the sound), [5] for percussive sounds, [6] for generic sounds. Other systems focus on describing sounds using the most perceptually significant characteristics (based on experimental results). For example [7] [8] [9] [10] propose systems based on perceptual features (often the musical instrument timbre) in order to allow application such as search-by-similarity or query-by-example. For these applications the underlying sound description is hidden to the user and only the results of the similarity search are given to him. This is because it is difficult to share a common language for sound description [11] outside the usual source/causal description. Therefore, a problem arises when dealing with abstract sounds, sound-effects, unnatural or synthetic sounds for which the source/cause is usually unknown or unrecognizable. Another approach must be used for these sounds.

A. Motivating applications

Being able to automatically describe sounds has immediate uses for the development of sound search-engines such as FindSounds.com, FreeSound.org or Ircam Sound Palette online. These search-engines are, however, limited to queries based on source/cause description, while providing search-by-similarity facilities.

In this paper, we study a sound description which is independent of the sound source/cause. The motivation is to extend the applicability of search-engines to sounds for which the cause is unknown, such as abstract sounds, sound-effects, unnatural or synthetic sounds.

A second motivation is to improve the usability of search-engines for sounds for which the cause is known but its description is not sufficiently informative to find sounds. Examples of this are the source descriptions of some environmental sounds. A sound referred to as “car” can indeed contain “car-door closing”, “car engine”, “car passing”, sounds which are completely different sounds. A “submachine gun” sound is often referred to as a “gun” with a specific trademark. Having the description that this sound is an iterative “gun” sound would be very useful. For this reason, sound designers, whose work is to find sounds to illustrate a given action, usually rely on a deeper acoustic description of the sound content. Unfortunately, this deeper description is usually personal to each sound designer who uses their own system. Sound designers also often use sounds from a completely different source to illustrate a target source because both have similar acoustic content. For example “stream” sounds are usually not done by recording a “stream” (which typically results in a white-noise sound) but by recording water in a bathroom. They also illustrate specific actions using sound with specific properties. For example, in order to illustrate an action with an increasing tension, one can use sounds with increasing dynamics and melody. In the opposite case, sounds with stable dynamics and melody are neutral and are therefore used for background ambiance.

In order to develop a system allowing automatic description of these sounds we need to

- choose an appropriate description of these sounds
- develop a system able to automatically index these sounds using the chosen description

The sound description we will use in this paper is the “morphological description” of sound object proposed by Schaeffer. We first review it in Section I-B. Existing audio features and classification algorithms that could be used to automatically

G. Peeters and E. Deruty are with the Sound Analysis/Synthesis Team ofIRCAM-CNRS STMS, 75004 Paris, France (e-mail: peeters@ircam.fr; deruty@ircam.fr).

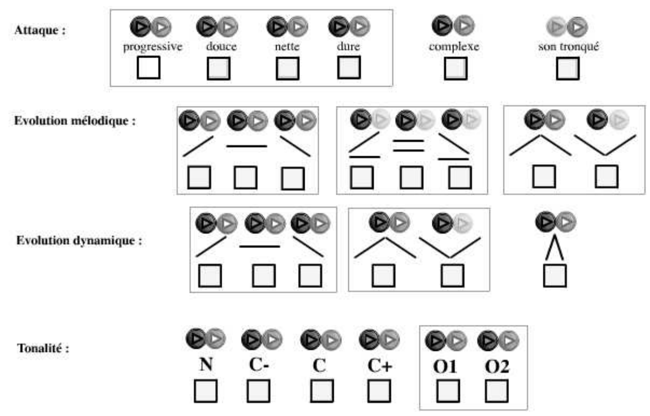

Fig. 1. Flash graphical interface for iconic representation of the main morphological profiles and sound descriptors.

index a sound in this description are then reviewed in Section I-C.

As an example of a final application using this system, we indicate in Fig. 1, the Flash-based interface of a sound search-engine for sound designers using this morphological descriptions. The various icons represent the various categories of the morphological descriptions. This interface allows the user to create a query based on the choice of specific categories: specific attack, complexity, melodic profile, dynamic profile, tonality. This work has been made in the framework of the Ecrins and Sample-Orchestrator project (with Digigram and Univers-Sons companies respectively).

B. Background on morphological sound description:

1) Schaeffer's morphological sound description: In "Traite des objets musicaux" [12] (later reviewed by [13]), Schaeffer proposes to describe sound using three points of view. The first one, named **"causal" listening**, is related to the sound recognition problem (when one tries to identify the sound source). The second, named **"semantic" listening**, aims at describing the meaning of a sound, the message the sound brings with it (hearing an alarm or a church-bell sound brings information). It is deeply related to the shared cultural knowledge. Finally the **"reduced" listening** describes the inherent characteristics of a sound independently of its cause and its meaning. The reduced listening leads to the concept of "sound object". A sound object is described using **morphological criteria**. Schaeffer distinguishes two kinds of morphology:

- the internal morphology, which describes the internal characteristics of a sound,
- the external morphology, which describes a sound object as being made of distinct elements, each having a distinctive form.

To distinguish between both he defines the concept of "unitary sound". A unitary sound contains only one event and cannot be further divided into independent segments, either in time (succession) or spectrum (superposition, polyphony).

Schaeffer proposes to describe sound using seven **morphological criteria**: the mass, the harmonic-timbre, the grain, the "allure", dynamic criteria, melodic profile and mass profile. These criteria can be grouped [14] into

- description of the sound matter: mass (description of the pitched nature of the sound), harmonic-timbre (dark, bright), grain (resonance, rubbing, iteration),
- description of the sound shape: dynamic criteria (impulse, cyclic), "allure" (amplitude of frequency modulation),
- variation criteria: melodic and mass profiles.

2) Modifications of Schaeffer's morphological sound description: Following Schaeffer works, there has been much discussion concerning the adequacy or not of Schaeffer criteria to describe generic sound, to verify their quality and pertinence. Some of the criteria, although very innovative (e.g. "grain", "allure" (rate), "profile") are very often subject to interrogations or confusions and have to be better circumscribed. Because of that, some authors have proposed modifications or additions to Schaeffers criteria [15] [16]. In the Ecrins project (Ircam, GRM, Digigram) [17], a set of criteria based on Schaeffers work has been established for the development of an online sound search-engine. The search-engine must use sound descriptions coming from automatic sound indexing. In this project, the morphological criteria (called morphological sphere) are divided into two sets of descriptors: main and complementary [18].

The **main descriptors** are: the duration, the dynamic profile (stable, ascending, or descending), the melodic profile (stable, up or down), the attack (long, medium, sharp), the pitch (either note pitch or area) and the spectral distribution (dark, medium, strident).

The **complementary descriptors** are the space (position and movement) and the texture (vibrato, tremolo, grain).

C. Background on automatic sound description:

Building a realistic sound search-engine application requires the ability to automatically extract the chosen sound description. In this part, we review existing audio features and classification algorithms which could be used to perform this task.

1) Audio features: An audio feature (sound descriptor) is a numerical value which describes a specific property of an audio signal. Typically, audio features are extracted by applying signal processing algorithms (such as FFT, Wavelet) to an audio signal. Depending on the audio-content (musical instrument sound, percussion, sound-effects, speech or music) and on the application (indexing or search-by-similarity) numerous audio features have been proposed such as the spectral centroid, Log-Attack-Time, Mel frequency cepstral coefficients etc. A list of the most commonly used audio features can be found in [19].

2) Modeling time: Audio features are usually extracted on a frame-basis: a value (a vector of values) is extracted every 20ms using a sliding analysis window of length around 40 to 80ms. We call these features "frame-based" or **"instantaneous" features** since they represent the content of the signal at a given "instant" of a signal: around the center of the frame. A sound is then represented by the succession of its instantaneous features over time. This notion of "succession" is however difficult to represent in a computer.

This is why the temporal ordering of the features is often represented using temporal derivatives: delta-features or acceleration-features. The features can also be summed up using their statistical moments over larger periods of time (by computing the mean and standard deviation of instantaneous features over a 500ms sliding-window) or by estimating a diagonal or multivariate auto-regressive model of the temporal evolution of the features [20]. Other models have also been proposed such as the use of the amplitude spectrum of the feature temporal evolution (named either dynamic [21], "penny" [22] or modulation spectrum [23] features). These features are often called **"texture window" features**.

When the temporal-modeling is applied directly over the whole file duration, we name the resulting features **"global" features** since they apply globally to the file and not to a specific time in the file (such as the mean and standard deviation of instantaneous features over the whole file duration).

Usually audio indexing problems are solved by computing instantaneous features, computing their corresponding "texture window" features and then applying pattern matching algorithms (such as Gaussian mixture models, Support Vector Machine). This approach is known as the "bag-of-frames" approach.

When the bag-of-frames approach is used, **late-integration algorithms** can be used in order to attribute a single class to the whole file from the class membership of each individual frame. When using a "bag-of-frames" approach in Section III, we will use a majority-vote among the class membership of each individual frame.

The notion of "succession" can also be represented using **time-dependent statistical models** such as Hidden Markov Models. In this case, a specific HMM models directly the belonging of all the frames of a file to a specific class.

D. Paper contribution and organization

There are a large number of papers related to sound indexing, presenting innovative features to describe the timbre or the harmonic content of the sound. However few of them deal with the problem of describing the shape of a sound, that is the shape of its features.

The goal of this paper is to propose new audio features to allow automatically indexing sounds as shapes. For this, we rely on a subset of the "sound object" description presented in Section I-B2. Among the presented descriptions, we focus on the descriptions related to the shape of the signal: the morphological descriptions. We consider the three following morphological descriptions, which are considered as three separated classification problems:

- Dynamic profiles. It is a problem with 5 classes: ascending, descending, ascending/descending, stable and impulsive.
- Melodic profiles. It is a problem with 5 classes: up, down, stable, up/down and down/up.
- Complex-iterative sounds. We first distinguish between non-iterative and iterative sounds. We then distinguish between the "grain" and "repetition" sounds inside the iterative sound class.

In Section II, we propose for each of the three problems new audio features that allow performing automatic indexing into the classes. We start by indicating the methodology used for the design of the audio features in Section II-A. We then propose specific features for the dynamic profiles in Section II-B, melodic profiles in II-C and complex-iterative sounds II-D. In Section II-E, we indicates audio features for the remaining descriptions of part I-B2.

We then evaluate the performances of the proposed audio features for automatic classification in Section III. For this we use the base-line classifier described in Section III-A. During the evaluation, we compare the results obtained using the proposed audio features with the ones obtained using standard audio features. We present the standard audio features in III-B. The results of the evaluation are presented for the three morphological descriptions in III-C, III-D and III-E.

We finally discuss the results in Section IV and present further work.

II. Audio features for morphological description

A. Methodology

Usually, audio classification systems work in an "a-posteriori" way: the systems proposed by [2] [6] [4] or [24] try a-posteriori to map (using statistical models) extracted audio features to the definition of a sound class. In this work, we propose to work the opposite way using a "prior" approach: we develop audio features corresponding directly to the classes of interest. Because the classes of interest are based on time-evolution, we integrate directly the notion of time in the definition and the design of the features[1] In order to do that, we need to understand the exact meaning of the morphological profiles in terms of audio content. We do this by using sets of audio files to illustrate each morphological profile[2].

1) Annotation:: The audio files have been selected by a sound designer based on their perceptual characteristics. Sounds that cannot be decomposed further or which decomposition would lead to segment-lengths shorter that the human ear integration time (around 40ms) have been classified as "unitary". For dynamic profiles, a sound has been qualified as "ascending" if its most important part (in terms of time duration) is perceived as having increasing loudness. Sounds with very slow and long attack (relative to the time duration of the sound) can therefore be qualified as "ascending". Likewise, sounds with very slow and long decay or release can be qualified as "descending". A sound is qualified as "stable" if no significant (in terms of time duration and amount of variation) variation of its loudness is perceived. A sound is qualified as "ascending/descending" if the sound is perceived as two parts, the first being perceived as with increasing

[1] A well known example of such a "prior" approach including time in the design of the features is the Log-Attack-Time which describes directly the length of the attack of a sound.

[2] Examples of sounds coming from these sets can be found at the following URL: http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ieeemorphological/.

loudness, the second one with decreasing loudness. The same rules have been applied for the melodic profiles. Note that the “ascending/descending”, “up/down” and “down/up” profiles are limit cases of the “unitary” sound definition and could have also been considered as the concatenation of two “unitary” sounds. However, since the two parts of the sounds used to illustrate these profiles were perceived as coming from the same process (continuity of loudness, pitch or timbre), they were considered here as “unitary” sounds.

2) File description:: All sounds are taken from the “Sound Ideas 6000” collection [25]. They are either real recordings (but recorded in clean conditions) or synthetic sounds. The audio files have a sampling rate of 44.1KHz and are monophonic. The shortest audio file duration is 0.0664s, the longest duration is 25.4898s.

B. Dynamic profiles

The description of a sound using dynamic profiles aims at qualifying a sound as belonging to one of the five following classes:

- ascending,
- descending,
- ascending/descending,
- stable,
- impulsive.

The sounds which are to be described are supposed to be unitary (i.e. cannot be further segmented). The profiles are illustrated by a set of 187 audio files taken from the “Sound Ideas 6000” collection.

1) Loudness:: According to this example set, the dynamic profiles are related to the perception of loudness. Therefore, in order to estimate the profiles, we first estimate the instantaneous loudness $l(t)$ from the signal. For this, the DFT of the signal is computed over time using short-term analysis. A filter simulating the mid-ear attenuation [26] is then applied to each DFT. The filtered DFT is then mapped to 24 Bark bands [27]. We note $E(z)$ the value of the energy inside the band z. We use an approximation of the specific loudness by neglecting terms of the expression acting only in specific cases (very weak signals) and by expressing it on a relative scale: $l'(z,t) = E(z,t)^{0.23}$. The loudness is the sum of the specific loudness: $l(t) = \sum_{z=1}^{Z} l'(z,t)$.

The loudness function over time is used to estimate the various dynamic morphological profiles.

2) Slope estimation:: The profiles “ascending”, “descending”, “ascending/descending” and “stable” are described by estimating the slopes of $l(t)$. We define t_M as the time which corresponds to the maximum value of the loudness over time. t_M is estimated from a smoothed version of $l(t)$. The smoothed version is obtained by filtering $l(t)$ with a 5th order F.I.R. low-pass filter with a 1Hz cut-off frequency[3]. As illustrated in Fig. 2, $l(t)$ is approximated using two slopes: one before t_M noted $S1$ and one after noted $S2$. The decay of natural sounds often follows an exponential law which can be expressed as $l(t) = A\exp(-\alpha(t-t_M))$,$t \geq t_M$ with A the value of $l(t)$ at the maximum position and $\alpha \geq 0$ the decay coefficient. We therefore express the loudness on a log-scale in order to estimate the first order polynomial approximation of the envelope.

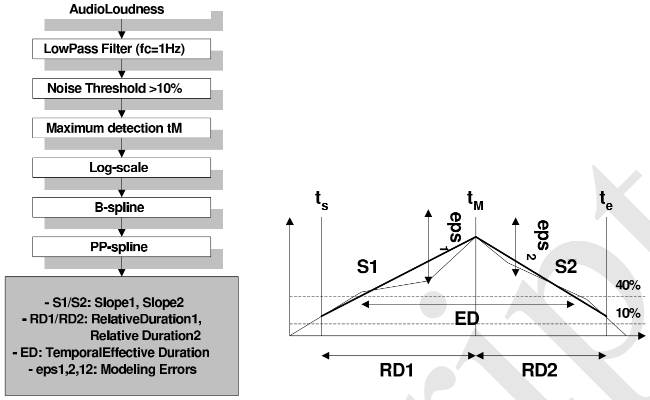

Fig. 2. Flowchart of the extraction algorithm of the audio features for the estimation of dynamic profiles.

3) Relative duration:: A small or large value of slope means nothing without the knowledge of the segment duration it describes. We define the relative-duration as the ratio of the duration of a segment to the total duration of the sound. We compute two relative-durations corresponding to the segments before and after t_M, noted $RD1$ and $RD2$ in the following. $RD1$ and $RD2$ are illustrated in Fig. 2.

4) Time normalization:: The dynamic profiles must be independent of the total duration of the sound (the loudness of a sound can increase over 1s or over 25s, it is still an “ascending” sound). For this, all the computations are done on a normalized time axis ranging from 0 to 1. As a consequence $RD2$ is now equal to $1 - RD1$.

5) B-spline approximation:: In order to obtain the slope corresponding to the dynamic profiles we want to approximate $l(t)$ by two first-order polynomials before and after t_M. However, this would not guarantee the continuity of the corresponding function at t_M. We therefore use a second-order B-spline to approximate $l(t)$ with knots at $(t_s, l(t_s))$, $(t_M, l(t_M))$ and $(t_e, l(t_e))$. t_s and t_e are the times corresponding to the first and last value of $l(t)$ above 10% of $l(t_M)$. Since the second-order B-spline is continuous at the 0th order, the resulting first-order polynomials before and after t_M are guaranteed to connect at t_M. In Fig. 3 we illustrate the extraction process on a real signal belonging to the “ascending/descending” dynamic profile. The B-spline approximation is then converted to its PP-spline form and the following set of features are derived from it (see Fig. 2): - S1: Slope of the first segment, - RD1: Relative Duration of the first segment, - S2: Slope of the second segment, - RD2: Relative Duration of the second segment.

6) Modeling error:: The adequacy of the two-slopes model to describe the time evolution of $l(t)$ is characterized using three modeling errors: $\epsilon_1 = \frac{\sum_{t=t_s}^{t_M}(\hat{l}(t)-l(t))^2}{\sum_{t=t_s}^{t_M}(l(t))^2}$ where $\hat{l}(t)$ is the

[3] Note that this filtering is applied to the loudness function $l(t)$ and not the audio signal. The 1Hz cut-off frequency was found to provide more consistent results over audio files. Using such a low frequency was possible because of the long duration of the “ascending” and “descending” segments. The loudness function is extracted from Short Term Fourier Transform using a Blackman analysis window of length 60ms with a 20ms hop size. It has therefore most of its energy below 19.58 Hz although its sampling rate is 50Hz.

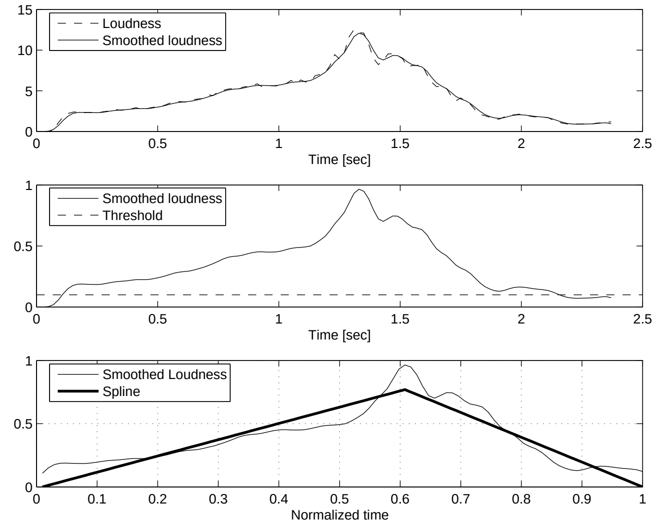

Fig. 3. Illustration of the estimation of dynamic profiles on a real signal from the class "ascending/descending": [TOP] loudness and smoothed loudness of the signal over time, [MIDDLE] 10% threshold applied to the smoothed loudness, [BOTTOM] smoothed loudness above the threshold and B-spline modeling.

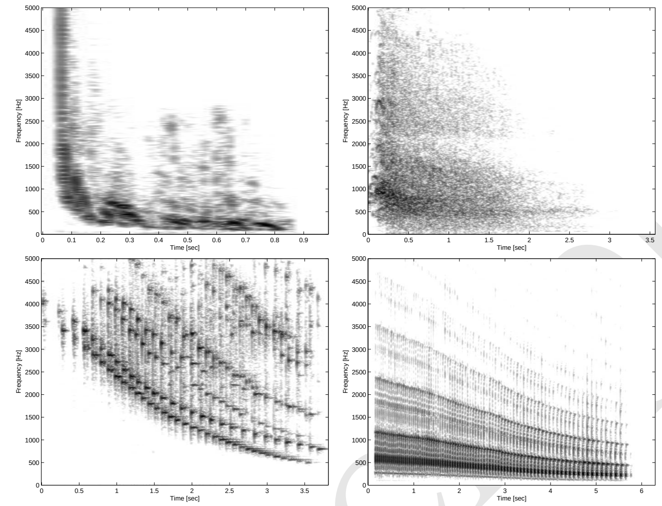

Fig. 4. Four signals (in spectrogram representation) belonging to the "down" melodic profile: [TOP-LEFT] discontinuous variation of the spectral envelope, [TOP-RIGHT] continuous variation of the spectral envelope, [BOTTOM-LEFT] succession of pitched notes, [BOTTOM-RIGHT] succession of events with decreasing resonances.

modeling of $l(t)$ obtained using the B-spline approximation. ϵ_2 and ϵ_{12} are computed in the same way on the intervals $[t_M, t_e]$ and $[t_s, t_e]$ respectively.

7) Effective duration: : The two-slope model allows representing the "ascending", "descending", "ascending/descending" profiles as well as the "stable" profile. The distinction between the "impulsive" profile and the other ones is done by computing the Temporal-Effective-Duration of the signal. The Temporal-Effective-Duration is defined as the duration over which $l(t)$ is above a given threshold (40% in our case), normalized by the total duration [19]. The Temporal-Effective-Duration is noted ED in the following. ED is illustrated in Fig. 2.

C. Melodic profiles

The description of sound using melodic profiles aims at qualifying a sound as belonging to one of the five following classes:

- up,
- down,
- stable,
- up/down,
- down/up.

For this description the input sounds are also supposed to be unitary. The profiles are illustrated by a set of 188 audio files taken from the "Sound Ideas 6000" collection [25].

While the dynamic profiles are clearly related to the perception of the loudness of the signal, the relationship between the melodic profiles and the signal content is not unique. Despite the shared perception of the melodic profiles, this perception comes either from a continuous modification of the pitch, a succession of separated pitched events, a continuous modulation of the spectral envelope, or a succession of discontinuous modulations of the spectral envelope (increasing or decreasing resonances). Other factors also play an important role in the perception of the profiles: the time extent over which the modulations occur and the intensity of the sound during the modulations. We illustrate this in Fig. 4. We represent four sounds belonging to the "down" melodic profile. For each of these sounds, the perception of the "down" profile is related to a different signal content modulation.

In order to describe the melodic profiles, we first tested the use of common spectral features: spectral centroid, spectral spread [19] or the perceptual features "sharpness" and "spread" [28]. The temporal evolution of these features is only weakly correlated to the melodic profiles. Tests of the applicability of sinusoidal modeling and pitch detection algorithms to sound-effects have also been performed. Unfortunately, because of the high variability of the sound-effects and the inadequacy of the sinusoidal or harmonic models to model sound-effects with noise-like sounds, these algorithms failed most of the time (highly discontinuous tracks or pitches). The spectrogram representation of the sounds of Fig. 4 provides a better understanding of this.

1) Tracking over-time of the most-excited filter:: The feature we propose for the description of the melodic profiles is based on the tracking over time of the most excited perceptual filter. This allows taking into account both pitch variation and resonant-frequency variation.

For this, we first compute the DFT of the signal over time using a short-term analysis. The DFT is then mapped to a set of 160 Mel[4] bands (triangular shape) acting as a set of perceptual filters. At each time t, we estimate the filter which has the maximum energy. In order to avoid discontinuities over time, the tracking over-time is performed using a Viterbi decoding algorithm. The Viterbi decoding algorithm takes as input: the initial probability that a filter is excited (set equal

[4] We have chosen to use Mel filters instead of Bark filters in order to have more freedom concerning the choice of the shape and the number of filters.

for all filters), the probability that a filter is excited at time t (represented by its energy), the probability to transit from one filter at time $t-1$ to another at time t (set as a Gaussian probability with standard-deviation equal to 5 filters). We note $f(t) \in \mathbb{N}$ describing the estimated path of filters over time t.

The melodic profiles could be correlated directly with $f(t)$. However, for several reasons we have decided to use a modification of this function:

1) The melodic profiles are said to be "up" or "down", not based on the amount of increase or decrease, but on the relative duration over which the change occurs and on the corresponding relative energy over which it increases or decreases.
2) Although the Viterbi decoding algorithm allows reducing octave errors, the function $f(t)$ still presents octave jumps. Therefore using $f(t)$ directly to match the profiles would mostly highlight octave jumps rather then the actual increase or decrease of the sound.

For these reasons, $f(t)$ is used to create another function defined as the cumulative integral over time of the weighted sign of the time-derivative of $f(t)$:

$$h(\tau) = \int_0^\tau e(t) \cdot sign\left(\frac{\delta f(t)}{\delta t}\right) \delta\tau \qquad (1)$$

The weighting factor $e(t)$ allows to emphasize the part of the signal with the highest intensity. $h(\tau)$ increases over time when the melodic profile increases and decreases when the profile decreases. In terms of implementation, the cumulative intergral is computed by a cumulative sum normalized by the total length of the considered signal. Also, only the part of the audio signal with energy above a given threshold is considered for the computation of $h(\tau)$. The threshold corresponds to 2% of the maximum energy value over time.

2) Slope estimation:: As for the dynamic profiles, we estimate the melodic profiles from the slopes derived from a B-spline approximation of $h(t)$. In the dynamic profiles, the knots of the B-spline were always positioned at the beginning t_s, maximum value t_M and ending time t_e of the function. For the melodic profiles, because of the presence of the "down/up" profile, the choice of a position t_M corresponding to the maximum value is problematic. Ideally, one would choose the position corresponding to the maximum value for the "down/up" profile and to the minimum value for the "up/down" profile. However, this would necessitate the estimation of the membership of "down/up" or "up/down" profile before doing the slope approximation. This estimation can be problematic for the "up", "down" and "stable" profile. Indeed, depending on the choice of maximum/minimum, a specific profile can be either represented by negative or positive $S1$, the same for $S2$. For example, "up" could be represented both by positive "S1" (with large duration) and negative "S2" (with short duration) or by negative "S1" (with short duration) and positive "S2" (with long duration). For these reasons, in the case of melodic profiles, the middle knot of the B-spline approximation, t_{middle}, is fixed and corresponds to the middle of the signal (the part of the signal above the 2% energy threshold).

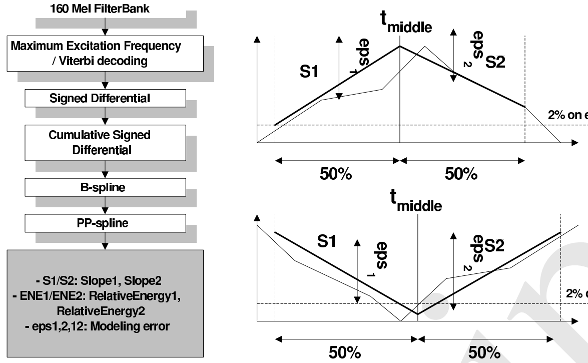

Fig. 5. Flowchart of the extraction algorithm of the audio features for the estimation of melodic profiles.

In Fig. 6, we illustrate the extraction process on a real signal belonging to the "down/up" melodic profile.

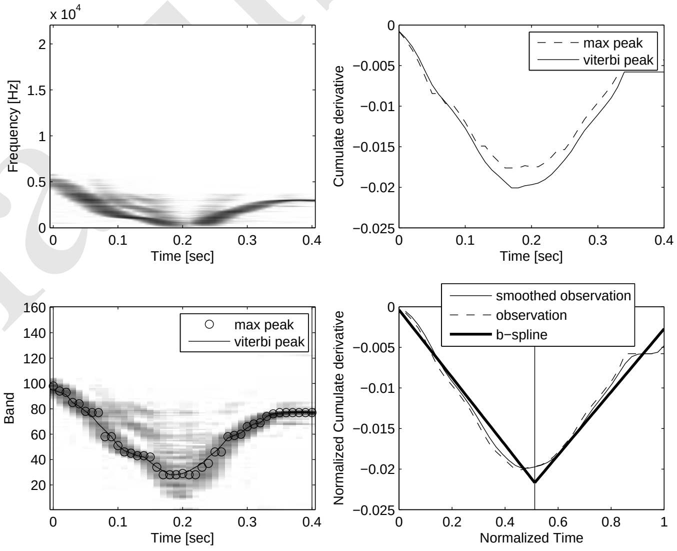

Fig. 6. Illustration of the estimation of melodic profiles on a real signal from the class "down/up": [TOP-LEFT] spectrogram of the signal, [BOTTOM-LEFT] conversion of the spectrogram to a 160 bands filter-bank, estimation of the filter with maximum excitation (circle) and Viterbi tracking (continuous line), [TOP-RIGHT] cumulative derivative $h(\tau)$ of the previous, [BOTTOM-RIGHT] B-spline approximation of the smoothed $h(\tau)$.

The B-spline approximation is then converted to its PP-spline form in order to derive the two slopes S1 and S2 (see Fig. 5). We also compute the relative energy contained in each of the two segments ENE1 and ENE2.

3) Modeling error:: The adequacy of the two-slopes model to describe the time evolution of $h(t)$ is characterized using three modeling errors: $\epsilon_1 = \frac{\sum_{t=t_s}^{t_{middle}} (\hat{h}(t)-h(t))^2}{\sum_{t=t_s}^{t_{middle}} (h(t))^2}$ where $\hat{h}(t)$ is the modeling of $h(t)$ obtained using the B-spline approximation. ϵ_2 and ϵ_{12} are computed in the same way on the intervals $[t_{middle}, t_e]$ and $[t_s, t_e]$ respectively.

The extraction process for melodic profile is summarized in Fig. 5.

D. *Complex-iterative sound description*

Dynamic and melodic profiles are descriptions of unitary sounds, i.e. description of sounds that cannot be further decomposed into sub-elements. Alternatively, complex sounds are composed of several elements. Iterative sounds are complex sounds defined by the repetition of a sound-element over time.

In this part, we propose features to distinguish between unitary and complex-iterative sounds. For complex-iterative sounds we propose a method to estimate their main periodicity which acts as the main characteristic to allow differentiating between "grain" (short period) and "repetition" (long period). For the "repetition" class, we propose an algorithm inspired by P-Sola analysis in order to segment the signal into repeated elements and allow further characterizing them in terms of dynamic and melodic profiles.

The complex-iterative sounds are illustrated by a set of 152 sounds taken from the "Sound Ideas 6000" collection.

Iterative sounds are defined by the repetition of a sound-element over time. Repetition of a sound-element can occur at the dynamic level, perceived pitch level or at the timbre level. This complicates the automatic detection of the repetition. Moreover several repetition cycles can occur at the same time for the same parameters (given a complex cycle such as the repetition of a rhythm pattern) or for various parameters (one dynamic cycle plus a different timbre cycle). Corresponding to these are methods for the automatic detection of repetition based on loudness, fundamental frequency or spectral envelope. Another complexity comes from the variation of the period of repetition over the sound duration or from disturbance from other perceived parameters.

In order to allow distinguishing between unitary and complex-iterative sounds and to allow distinguishing between "grain" and "repetition" we propose the following characteristics:

- The **periodicity**/amount of repetition: allows distinguishing between iterative sounds and non-iterative sounds.
- The **period** of the cycle: allows distinguishing between "grain" (short period) and "repetition" (long period).

In the case of "repetition", we also propose an algorithm for segmenting the signal into events and to allow to further describe them in terms of dynamic and melodic profiles.

1) Periodicity and period of the cycle:: Mel Frequency Cepstral Coefficients (MFCCs) are first extracted from the signal with a 50Hz sampling rate. The 0th order coefficient of the MFCC represents the global energy of the spectrum, hence that of the entire signal. This representation then takes into account both energy variations and spectral variations. We use 12 MFCCs derived from a 40 triangular-shape Mel bands analysis. We denote $\underline{o}(t)$ as the vector of MFCCs at time t. The similarity matrix [29], which represents the similarity between each pair of vectors, is then computed using an Euclidean distance. We denote it $\underline{\underline{S}}(t_i, t_j) = d(\underline{o}(t_i), \underline{o}(t_j))$. $\underline{\underline{S}}(t_i, t_j)$ is then converted to the corresponding lag-matrix [30] $\underline{\underline{L}}(l_{ij}, t_j)$ with $l_{ij} = t_j - t_i$. We define an AudioSimilarity function as the normalized sum of the lag-matrix (which is a lower-triangular matrix) over the time axis:

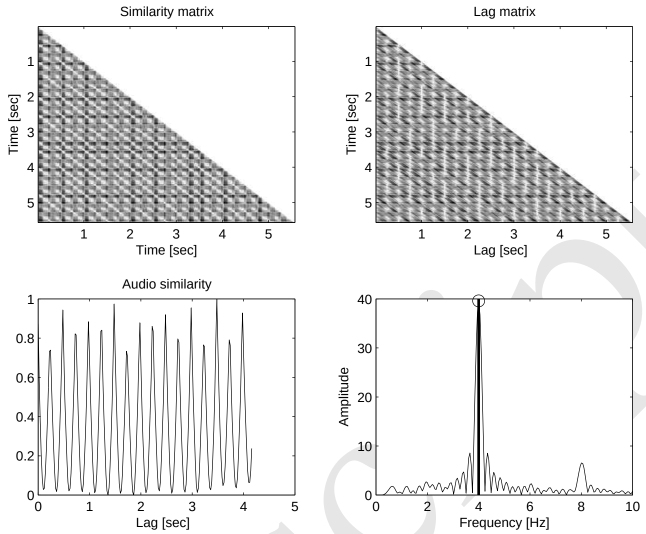

Fig. 7. Illustration of the estimation of the description of complex-iterative sounds: [TOP-LEFT] MFCC similarity matrix, [TOP-RIGHT] corresponding lag-matrix, [BOTTOM-LEFT] audio similarity function, [BOTTOM-RIGHT] amplitude spectrum of the audio similarity function and estimation of the frequency of the cycle (circle, thick vertical continuous line).

$$a(l) = \frac{1}{J - l + 1} \sum_{j=l}^{J} \underline{\underline{L}}(l, t_j) \tag{2}$$

where J is the total number of discrete times t_j. This AudioSimilarity function expresses the amount of energy and timbre similarity of the sounds for a specific lag l.

The AudioSimilarity function $a(l)$ is then used to compute the periodicity and period of the cycle. For this, we compute the amplitude spectrum of the AudioSimilarity function $a(l)$ and estimate its maximum peak within the range [0.1, 20] Hz. We note M the amplitude value of the maximum peak. We then choose the lowest-frequency peak which has an amplitude $\geq 0.5M$ as the peak representing the frequency $1/T_0$ of the cycle. The periodicity (amount of repetition) is given by the value of the normalized auto-correlation function of $a(l)$ at the lag corresponding to the period of the cycle T_0.

2) Localization and characterization of one of the repeated elements:: Given the estimated period T_0 of the cycle, the localization of the repeated elements is done by a method previously developed for P-Sola pitch-periods localization [31]. For this, we define a vector of cycle instants $\underline{T_\tau(t)} = \sum_k \delta(t - \tau - kT_0)$ ($\underline{T}$ is a comb-filter starting at time τ with periodicity T_0). We define $e(t)$ as the energy function (RMS value) of the signal. The local-minima of $e(t)$ around the values of $\underline{T}$ are detected. We compute the sum $E(\tau)$ of the value of $e(t)$ at these local-minima positions. The process is repeated for various τ values. The value τ leading to the minimum value of $E(\tau)$ defines the vector which gives the best time locations for a segmentation into repeated elements.

This process is illustrated in Fig. 7 and 8 for a real signal.

Given the estimated location of the repeated element, we isolate one of the elements in the middle of the sound and characterize its acoustic content in terms of dynamic and

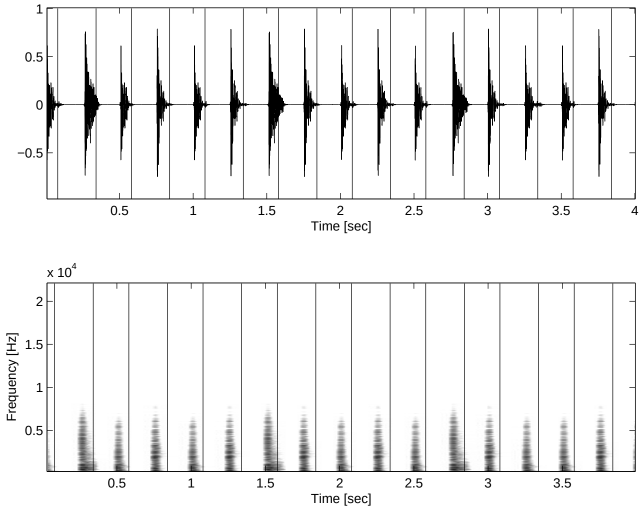

Fig. 8. Illustration of the estimation of the description of complex-iterative sounds: [TOP] signal waveform and segmentation estimated using the period of cycle and P-Sola algorithm, [BOTTOM] corresponding spectrogram and segmentation.

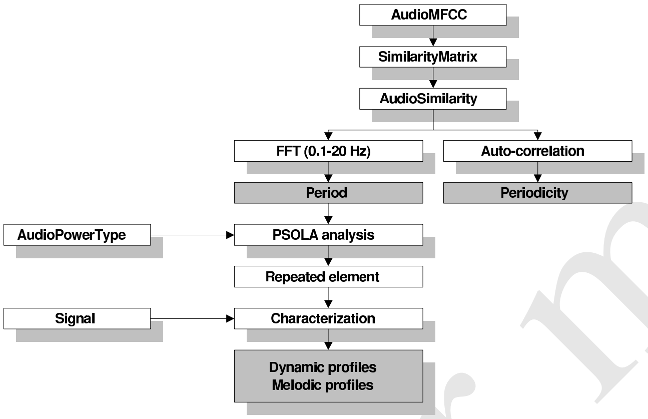

Fig. 9. Flowchart of the extraction algorithm of the audio features for the description of complex-iterative sounds.

melodic profiles.

The flowchart of the extraction process is illustrated into Fig. 9.

E. Remaining description

The remaining descriptions of sound objects presented in Section I-B2 (duration, attack, pitch and spectral distribution) are not discussed in this paper since they do not involve modeling time. These descriptions can be obtained using the audio features described in [19] and were discussed in previous works such as [14]. For example, the duration can be obtained using the Temporal-Effective-Duration feature, the description of the attack using the Log-Attack-Time (an efficient method to estimate it has been proposed in [19]), the pitch using numerous existing pitch estimation algorithms [32] [33] [34] [35], the spectral distribution using the perceptual features spectral centroid and spectral spread.

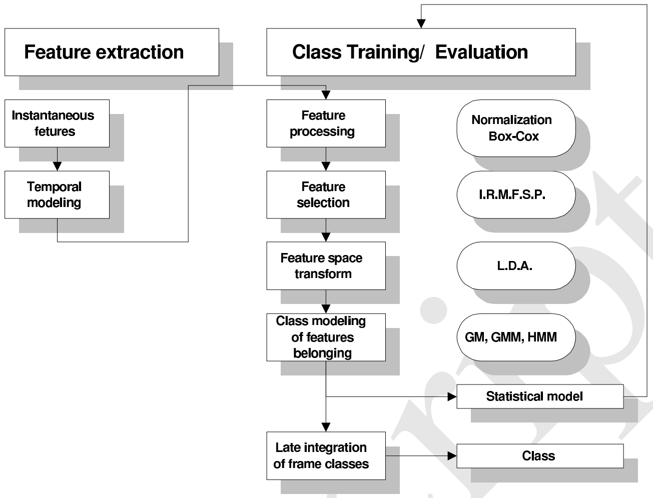

Fig. 10. Base-line automatic classifier of [24].

III. Evaluation

In this part, we evaluate the performances of the proposed audio features for automatic classification into the dynamic profile classes, melodic profile classes and for the discrimination between iterative and non-iterative sounds. Inside the class of iterative sounds, we also test the ability of the features to estimate the period of the sounds in order to separate the classes “grain” (short period) and “repetition” (long period). For the classification tasks, we will use the generic automatic classification system of [24]. We present this system in Section III-A. For each classification task, we compare the results obtained with the proposed audio features to the results obtained using standard audio features using the same classification system. We present the standard audio features we have used in Section III-B. We then discuss the three classification tasks in Sections III-C, III-D and III-E.

A. Base-line automatic sound description system

The system we use to perform automatic classification is the one described in [24]. It has been developed to solve automatically a large set of indexing problems[5]. The flowchart of this system is indicated into Fig. 10.

The system takes as input a set of **audio feature vectors** $f_{c,i,t}(k)$, where c represents belonging to a class, i belonging to a specific segment or audio file, t the time index in this segment/file and k the dimension in the feature vector (such as the k^{th} MFCC coefficient). In the case of “global” features, t is omitted since the features refer directly to the whole duration of the file.

An **automatic feature-selection (AFS)** algorithm is then used to find the best features (the best dimensions k) to

[5] This system has been shown to be able to achieve good indexing results for a variety of applications: - when applied to the problem of “musical instrument sample” indexing [3], it was qualified by [36] as “probably a fair representative of the current state of the art in instrument classification”. - when applied to music indexing: it has won the Audio Mood classification, ranked 2nd in the Audio Genre Classification (Genre-Latin) and Audio Classical Composer identification tasks in the MIREX08 contest [37]. We therefore think this system is a good base-line system.

discriminate between the various classes. We have used our Inertia Ratio Maximization with Feature Space Projection (IRMFSP) algorithm. In [38], we positively compared the IRMFSP algorithm to the most-used CFS [39] algorithm. The IRMFSP algorithm is an iterative algorithm which selects one feature at a time. The selection is based on the feature with the largest Fisher ratio. It then performs a Grahm-Schmidt orthogonalization of the whole feature space on the selected feature. This guarantees that the remaining features are no longer correlated with the selected features. The selection process is then repeated until the Fisher ratio passes under a specific threshold.

Classification models based on Gaussian distribution makes the underlying assumption that the modeled data (the various dimensions k) follow a Gaussian probability density function (pdf). However, this is rarely the case. Therefore a non-linear transformation, known as the **"Box-Cox (BC)"** transformation [40] is applied to each feature k individually in order to make its pdf fit as much as possible to a Gaussian pdf

Linear Discriminant Analysis (LDA) [41] allows finding a linear combination among features in order to maximize the discrimination between the classes. LDA is applied to the features in order to reduce the dimensionality of the feature space while improving class separation.

Class modeling can then be achieved using the following model: - GM: multi-dimensional Gaussian model (with diagonal or full covariance matrix), - GMM: multi-dimensional Gaussian mixture model (with a varying number of components Q), - HMM: Hidden-Markov-Model (with a varying numbers of states S). The HMM can be used to model "instantaneous" or "texture-window" features. However, it cannot be applied to "global" features since in this case there is no time-evolution to model. HMM provides directly the class corresponding to the whole audio-file. When using GM or GMM to model "instantaneous" or "texture-window" features, we apply a "late-integration" algorithm to find the best single class explaining the whole audio-file. For this, we use a majority-vote among the class membership of each individual frame of a given file.

In the following, we indicate for each experiment the best parameters of the system: use of Automatic Feature Selection (AFS) or not, use of the Box-Cox transform (BC) or not, use of Linear Discriminant Analysis (LDA) or not, statistical model used (GM, GMM or HMM) and its parameters (Q, S and type of the covariance matrices).

The tests are performed using N-fold cross-validation. We have used a value of $N = 10$ for all cases except for the melodic profiles where a value of $N = 6$ was used in order to guarantee source independence between training and test sets.

In the following we will also use the Partial Decision Tree (PART) algorithm [42] in order to understand the values of the features specific to each class. We have used the Weka [43] implementation of the PART algorithm.

B. Base-line audio features

In the following, we will compare the results obtained using the proposed audio features to the results obtained using standard-audio features. We have used four sets of standard audio features representing different audio characteristics:

- Description of the spectral-shape: Mel-Frequency-Cepstral-Coefficients (13 coefficients including a DC component using 40 triangular-shape Mel-bands) [44].
- Description of the harmonic/noise content: Spectral-Flatness Measure (SFM) (4 coefficients representing the frequency bands [250,500] [500,1000] [1000,2000] [2000,4000] Hz) and Spectral-Crest-Measure (SCM) (4 coefficients) [45] [46].
- Description of the shape and the harmonic/noise content: Spectral Peak (4 coefficients representing the same 4 frequency bands), Spectral Valley (4 coefficients) and Spectral Contrast (4 coefficients) [47]
- Description of the harmonic content: 12 dimensional Pitch-Class-Profile (PCP) also named Chroma [48] [49].

We also estimate the delta and acceleration coefficients of each feature (obtained by derivations of the local polynomial approximation of the time trajectory on 5 points). We have used a 40ms analysis window with a 20ms hop size. The type of analysis window varies according to the feature extracted.

The use of an automatic feature selection algorithm will allow us to find the best sub-set of features for each classification task.

Considering the presence of very short sounds in the test-set (66ms) it was not possible to use the "texture-window" features for the experiments. We therefore only study the two following modelings: - direct use of "instantaneous" features, and - use of "global" features with mean and standard deviation temporal modeling (the mean and standard deviation are computed over the whole file duration).

C. Dynamic profiles

The proposed features for dynamic profile estimation have been evaluated on a test-set of 187 audio files (26 ascending, 68 descending, 24 ascending/descending, 37 stable, 32 impulsive) taken from the "Sound Ideas 6000" collection [25].

In Fig. 11 we represent the distribution of the 5 proposed features ($RD1$, $S1$, $RD2$, $S2$ and ED) for each class. In this figure we see that "impulsive" sounds are characterized by a small value of ED, "ascending" ones by a large $RD1$, "descending" ones by a large $RD2$, "ascending/descending" have almost equal values of $RD1$ and $RD2$, "stable" by small $S1$ and $S2$. We also represent the distributions of the three modeling errors (ϵ_1, ϵ_2 and ϵ_{12}).

We now test the applicability of the proposed features to perform automatic classification of a sound into the 5 dynamic profiles. For this, we use the proposed audio features as input to the classification system presented in Section III-A. The best results are obtained with the following configuration of the classifier: no AFS, no BC, no LDA, GM with diagonal covariance-matrices. With this configuration and a 10-folds cross-validation, the mean-recall (mean over the N-folds of the mean-over-class recall[6]) is **97%**. We indicate in Table I,

[6] Among the recall, precision and F-measure, only the recall does not dependent on the distribution of the test-set. For this reason, in this paper, we use the recall to measure the performances.

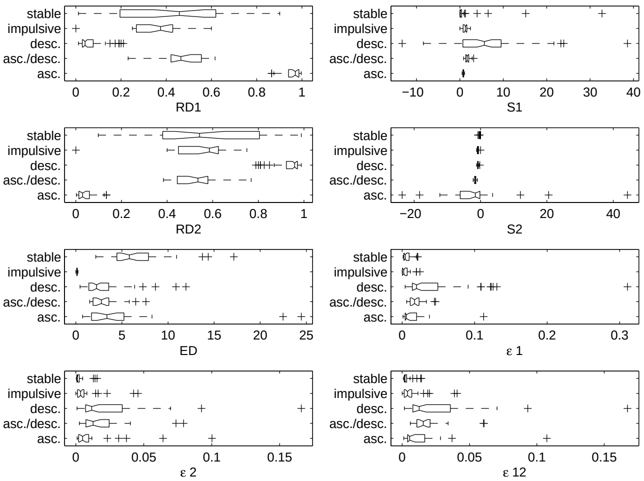

Fig. 11. Distribution (presented in the form of a box and whisker statistical plot) of the features $RD1$, $S1$, $RD2$, $S2$, ED and the three modeling errors ϵ_1, ϵ_2 and ϵ_{12} for the 187 sounds of the dynamic profiles test set.

TABLE I
CONFUSION MATRIX OF CLASSIFICATION INTO DYNAMIC PROFILES USING THE PROPOSED AUDIO FEATURES.

Real/Found	asc.	desc.	asc./ desc.	stable	impulsive	Items	Recall
asc.	26					26	100,0%
desc.		67		1		68	98,5%
asc./ desc.		1	23			24	95,8%
stable	1	2	1	33		37	89,2%
impulsive					32	32	100,0%
Items	27	70	24	34	32		**96,7%**
Precision	96,3%	95,7%	95,8%	97,1%	100,0%	**97,0%**	

the corresponding confusion matrix. As one can see the largest confusion occurs for the sounds from the “stable” class. This can be understood by the fact that “stable” is the limit case of “ascending”, “descending” and “ascending/descending”. In Table II we indicate the simple and intuitive set of rules found by the PART algorithm.

We now test the classification performances using the standard audio features presented in Section III-B with the same classification system. Using standard audio features in their instantaneous form, the best results are obtained with the following configuration of the classifier: AFS, no BC, no LDA, HMM with S=3, Q=1 and diagonal covariance matrices. The mean recall is then **59%**. Using standard audio features in their global modeling form, the best results are obtained with the following configuration of the classifier: AFS, BC, LDA, GMM with Q=3 and diagonal covariance matrices. The mean recall is then **76%**. It is interesting to note that the first selected feature using the AFS is the standard deviation of MFCC-0 (the DC-component) which corresponds to the definition of dynamic profiles.

D. *Melodic profiles*

The proposed features for melodic profiles estimation have been evaluated on a test-set of 188 audio files (71 up, 56 down, 32 stable, 23 up/down, 6 down/up) taken from the “Sound Ideas 6000” collection [25].

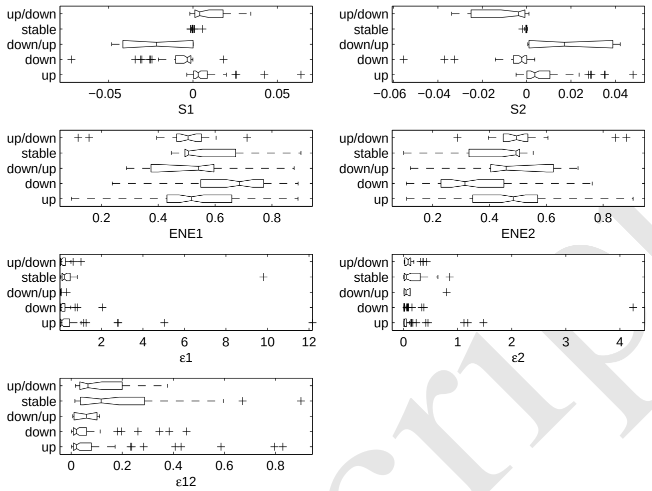

Fig. 12. Distribution (presented in the form of a box and whisker statistical plot) of the features S1, S2, ENE1, ENE2 and the three modeling errors ϵ_1, ϵ_2 and ϵ_{12} for the 188 sounds of the melodic profiles test set.

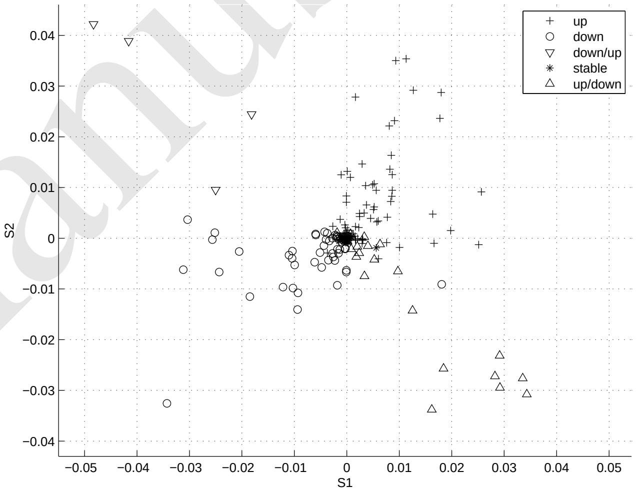

Fig. 13. Distribution of the features (presented in the form of a two-dimensional space) S1, S2 for the 5 classes of the melodic profiles test set

In Fig. 12 we represent the distribution of the 4 proposed features ($S1$, $S2$, $ENE1$ and $ENE2$) and the 3 modeling errors (ϵ_1, ϵ_2 and ϵ_{12}) for each class. As for the dynamic profiles, a clear trend of the features over the classes can be observed: $S1$ and $S2$ are large and positive for the “up” class, $S1$ and $S2$ are large and negative for the “down” class, $S1$ is positive and $S2$ negative and both are large for the “up/down” class, $S1$ is negative and $S2$ positive and both are large for the “down/up” class, $S1$ and $S2$ are small for the “stable” class. For a better visualization of $S1$ and $S2$, we represent in Fig. 13 the values obtained for each class in the $S1$ $S2$ two-dimensional space.

As for the dynamic profiles, we test the applicability of the proposed audio features to perform automatic classification of a sound into the 5 melodic profiles. The best results are obtained with the following configuration of the

TABLE II
RULES FOR AUTOMATIC CLASSIFICATION INTO DYNAMIC PROFILES OBTAINED USING THE PART ALGORITHM.

desc.	stable	asc.	impulsive	asc./ desc.
ED > 0.24	ED > 0.46	ED > 0.46	ED <= 0.82	S1 <= 3.824285
RD1 <= 0.210526	RD1 <= 0.73991	RD1 > 0.73991		S2 > -2.602062
S2 <= -0.198982	S2 > -0.863574			

TABLE III
CONFUSION MATRIX OF CLASSIFICATION INTO MELODIC PROFILES USING THE PROPOSED AUDIO FEATURES.

Real/Found	up	down	down/up	stable	up/down	Items	Recall
up	49	1	4	8	9	71	69,0%
down	0	41	4	7	4	56	73,2%
down/up	1	0	4	1	0	6	66,7%
stabe	1	3	0	27	1	32	84,4%
up/down	1	2	0	3	17	23	73,9%
Items	52	47	12	46	31		**73,4%**
Precision	94,2%	87,2%	33,3%	58,7%	54,8%	**65,7%**	

classifier: AFS, BC, no LDA and a GMM with Q=3 and diagonal covariance-matrices. The two selected features are “S1” and “S2”. With this configuration and a 6-folds cross-validation, the mean recall is **73%**. We indicate in Table III, the corresponding confusion matrix. The largest confusions are obtained between the classes “up” and “stable”, “down” and “stable”, “stable” and “down”. There is no confusion between the classes “up/down” and “down/up”. The lowest Precision occurs for the classes “down/up”, “stable” and “up/down”. This is a direct consequence of the unbalancing of the test-set.

The best result obtained with standard feature in instantaneous form is **29%** (using AFS, no BC, LDA, a GMM with Q=2 and diagonal covariance-matrices, and late-integration algorithm). Note that in the present case, the HMM did not produce the best results. The best result obtained with standard feature in global modeling form is **48%** (using AFS, BC, LDA and a GMM with Q=3 and diagonal covariance matrices).

E. Complex-iterative sound description

Classification into iterative and non-iterative sounds: We first test the applicability of the Periodicity (amount of repetition) feature to discriminate between the iterative and non-iterative sounds. For this we consider the 152 items of the iterative test-set as belonging to the iterative class, and all the items of the dynamic and melodic profiles test-set (unitary sounds) as belonging to the non-iterative class (375 items).

In Fig. 14, we represent the distribution of the features T_0 (period of the cycle) and Periodicity for the two classes. A clear separation between the two classes is visible for the Periodicity feature. Using only the Periodicity feature, we obtain the following mean recall (using no BC and a GM): **82%** (R_{iter}=85%, $R_{noniter}$=79%). We indicate the simple and intuitive rule obtained with the PART algorithm: $Periodicity \leq 0.7475$, $Periodicity > 0.7475$ (iteratif).

The best result obtained with standard features in instantaneous form is **66%** (using AFS, no BC, no-LDA and a HMM with S=3, Q=1 and diagonal covariance matrices). The best result obtained with standard features in global temporal modeling form is **82%** (using AFS, BC, LDA and a GMM with Q=3 and diagonal covariance matrices). This result is the same as the one obtained with the proposed Periodicity feature. Note however that in the Periodicity feature case, we only use a single feature and a very simple statistical model (simple GM).

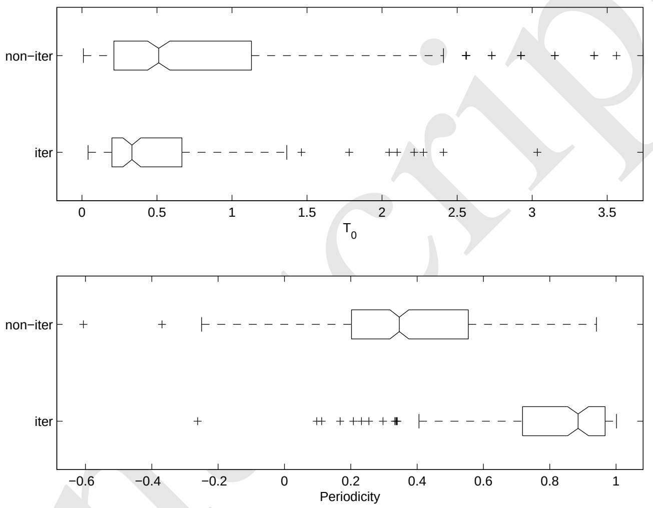

Fig. 14. Distribution (presented in the form of a box and whisker statistical plot) of the features T_0 and Periodicity for the 527 sounds of the non-iterative/iterative sounds test set.

Period estimation: We now evaluate the quality of the estimated period T_0 of the cycle for the iterative sounds. For this, only the sounds from the iterative test set having a single non-variable period over time are considered. This test-set of 67 iterative sounds has been manually annotated into cycles by one of the authors. To measure the quality of the estimated frequency $f_0 = 1/T_0$ of the cycle, we define two measurements: “Accuracy 1” is the percentage of frequency estimates within 4% of the ground-truth frequency. “Accuracy 2” is the percentage of frequency estimates within 4% of either the ground-truth frequency, 1/2 or 2 the ground-truth frequency. “Accuracy 2” allows taking into account octave errors in the estimation. We have obtained the following results using the proposed algorithm: Accuracy 1= **82.09%**, Accuracy 2= **89.55%**.

In Fig. 15 and 16, we present detailed results of the evaluation. We define r as the ratio between the estimated frequency and the ground-truth frequency. In Fig. 15, we represent the histogram of the values r in log-scale ($\log_2$) for all instances of the test-set. The vertical lines represent the values of r corresponding to usual frequency confusions:

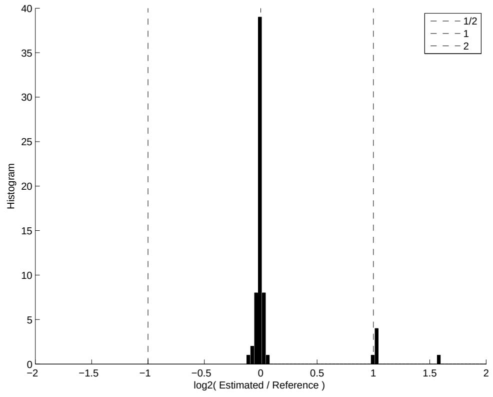

Fig. 15. Histogram of the ratio in log-scale between estimated and ground-truth frequency of the cycle.

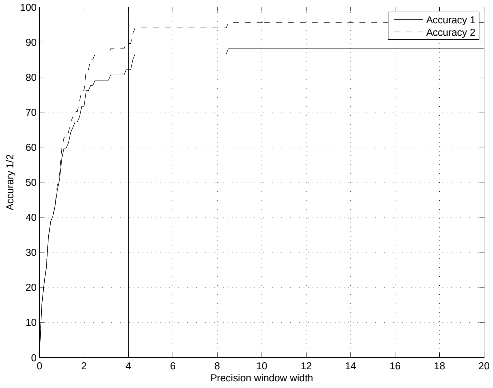

Fig. 16. Accuracy 1/2 versus precision window width (in % of correct frequency) for the iterative sound test-set.

1/2 (line at -1) and 2 (line at 1). The histogram indicates that all octave errors are up. There are no down-octave errors. In Fig. 16, we represent the influence of the precision window width on the recognition rate. The vertical line represents the precision window width of 4% used for the results. According to this figure, increasing the width until 8.5% would allow to increase both Accuracies up to 88.06% and 95.52%. After 8.5% the Accuracies remain constant, which means that the remaining estimation errors are “gross” errors.

IV. Conclusion

In this paper, we have studied the automatic indexing of sound samples using the morphological descriptions proposed by Schaeffer. Three morphological descriptions have been considered: dynamic profiles, melodic profiles, and complex-iterative sound description. For each description, a set of audio features have been proposed.

The dynamic profiles are estimated by modeling the loudness function over time by a second-order B-spline. The two slopes (S1, S2) and relative durations (RD1, RD2) derived from it, combined with the Temporal-Effective-Duration feature and modeling errors ($\epsilon_1, \epsilon_2, \epsilon_{12}$) allows a good classification (97%) into the 5 considered profiles. For comparison the best result obtained using standard audio features (MFCC, SFM/SCM, SP/SV/SC and PCP) is 76%.

The melodic profiles are estimated by tracking over time the perceptual filter which has the maximum excitation. From this track, we derive a function (the cumulative integral over time of the weighted sign of the time-derivative of the filter number) which is used to map the signal to the 5 considered profiles using again a second-order B-spline approximation. Given the complexity of the melodic profile estimation (multiple underlying criteria), the obtained results (73%) are judged good. The largest confusions occur between the “stable”, “up” and “down” which can be explained considering that the limit case of “up” is “stable” and the one of “down” is also “stable”. For comparison, the best result obtained using standard audio features is 48%.

Finally, we have studied the description of complex-iterative sounds. We have proposed an algorithm to estimate the Periodicity and the period of the cycle of the sound based on an AudioSimilarity function derived from an MFCC similarity matrix. This algorithm allows discriminating the non-iterative and iterative classes at more than 82%. However, similar results can be obtained using standard audio features in global temporal modeling form. The precision of the estimated period of the cycle is around 82% (90% if we consider octave errors as correct). The automatic estimation of the period cycle allows discriminating between the “grain” and “repetitions” classes.

As a conclusion, except for the discrimination between non-iterative and iterative classes, the proposed audio features seems to better catch the characteristics involved in the morphological dynamic and melodic profiles. The disappointing results obtained by modeling the standard audio features using HMM can be explained by the large variations of temporal extent of the profiles (some sounds are ascending over 1s, other ones over 25s). This variation did not allow the transition matrix of the HMM to catch the variations involved in the profiles.

The remaining description of sound objects (duration, description of attack, pitch and spectral distribution) were not discussed here since they can easily be achieved using previously existing works.

The main emphasis of this paper has been on the creation of dedicated audio features to solve complex indexing problems. We have used a “supervised” feature design approach using illustrative sound examples. Both the dynamic and melodic profiles were described by first extracting a signal observation and then forcing its temporal evolution to match a second-order B-spline model. While this approach may seem restrictive, it is important to notice that the most important feature coming from perceptual experiments [50], the attack time, is best described by the Log-Attack-Time feature, the design of which follows a similar approach to the one used here.

In this work, instead of relying on the results of perceptual experiments on sound listening from which descriptions are derived, we relied on prior descriptions (coming from Schaeffer proposals) and the knowledge/skill of a sound designer to exemplify these descriptions. This reverse-order is promising since it allows providing descriptions complementary to the ones found by experiments. However, further work should concentrate on validating this approach by performing posterior perceptual experiments, the goal of which will be to compare the proposed description to the ones obtained, on the same sounds, by experiments.

The features proposed in this work were created and used for morphological description of the sound. However, they can also be used for the usual "causal"/source description. Therefore, we believe that using these features it is possible to connect both approaches. Further work will concentrate on that.

Acknowledgment

The author would like to thank the two anonymous reviewers for their fruitful comments which help in the improvement of this paper. They would also like to thank Leigh for the careful English revision of the paper. Part of this work was conducted in the context of the Ecrins PRIAMM French Project, CUIDADO IST European Project, Sample Orchestrator ANR/RIAM French Project and Quaero Oseo French Project.

References

[1] K. Martin, "Sound source recognition: a theory and computational model," PHD Thesis, MIT, 1999.

[2] A. Eronen, "Comparison of features for musical instrument recognition," in *Proc. of IEEE WASPAA*, New Paltz, NY, USA, 2001.

[3] G. Peeters, "Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization," in *Proc. of AES 115th Convention*, New York, USA, 2003.

[4] S. Essid, "Classification automatique des signaux audio-frquences: reconnaissance des instruments de musique," PHD Thesis, Telecom ParisTech, 2005.

[5] P. Herrera, A. Dehamel, and F. Gouyon, "Automatic labeling of unpitched percussion sounds," in *Proc. of AES 114th Convention*, Amsterdam, The Nederlands, 2003.

[6] M. Casey, "General sound similarity and sound recognition tools," in *Introduction to MPEG-7 : Multimedia Content Description Language*, B. Manjunath, P. Salembier, and T. Sikora, Eds. Wiley Europe, 2002.

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

[8] N. Misdariis, B. Smith, D. Pressnitzer, P. Susini, and S. McAdams, "Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres," in *Proc. of 135th Meet. Ac. Soc. of America / 16th Int. Cong. on Acoustics*, Seattle, Washington, USA, 1998.

[9] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of mpeg-7," in *Proc. of ICMC*, Berlin, Germany, 2000, pp. 166–169.

[10] Comparisonics, "http://www.findsounds.com/," 2008.

[11] A. Faure, "Des sons aux mots : Comment parle-t-on du timbre musical ?" PHD Thesis, Ecoles des Hautes Etudes en Sciences Sociales, 2000.

[12] P. Schaeffer, *Trait des objets musicaux*. Paris: Seuil, 1966.

[13] M. Chion, *Guide des objets sonores*. Paris: Buchet/Chastel, 1983.

[14] J. Ricard and P. Herrera, "Morphological sound description: Computational model and usability evaluation," in *Proc. of AES 116th Convention*, Berlin, Germany, 2004.

[15] C. Olivier, "La recherche intelligente de sons," Master Thesis, Univ. de Provence, France, 2006.

[16] R. Leblanc, "Elaboration d'un systeme de classification pour sons figuratifs non instrumentaux," DESS Thesis, Universite Pierre et Marie Curie, Paris 6, 2000.

[17] P. Mullon, Y. Geslin, and M. Jacob, "Ecrins: an audio-content description environment for sound samples," in *Proc. of ICMC*, Goteborg, Sweden, 2002.

[18] E. Deruty, "Ecrins report: Descripteurs morphologiques / sons essentiels," Ircam, Ecrins Project Report, 2001.

[19] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project." Ircam, Cuidado Project Report, 2004.

[20] A. Meng, P. Ahrendt, J. Larsen, and L. Hansen, "Temporal feature integration for music genre classification," *IEEE Trans on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1654–1664, 2007.

[21] G. Peeters and X. Rodet, "Music structure discovering using dynamic audio features for audio summary generation: Sequence and state approach," in *Proc. of CBMI (Int. Workshop on Content-Based Multimedia Indexing)*, Rennes, France, 2003, pp. 207–214.

[22] B. Whitman and D. Ellis, "Automatic record reviews," in *Proc. of ISMIR*, Barcelona, Spain, 2004.

[23] S. Greenberg and B. E. D. Kingsburyy, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. of IEEE ICASSP*, vol. 3, Munich, Germany, 1997, pp. 1647–1650.

[24] G. Peeters, "A generic system for audio indexing: application to speech/music segmentation and music genre," in *Proc. of DAFX*, Bordeaux, France, 2007.

[25] Sound-Ideas, "The series 6000 "the general" sound effects library," 1992.

[26] B. Moore, B. Glasberg, and T. Baer, "A model for the prediction of thresholds loudness and partial loudness," *J. Audio Eng. Soc.*, vol. 45, pp. 224–240, 1997.

[27] E. Zwicker and E. Terhardt, "Analytical expression for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, vol. 68, pp. 1523–1525, 1980.

[28] E. Zwicker, "Procedure for calculating loudness of temporally variable sounds," *J. Acoust. Soc. Am.*, vol. 62, no. 3, pp. 675–682, 1977.

[29] J. Foote, "Visualizing music and audio using self-similarity," in *Proc. of ACM Int. Conf. on Multimedia*, Orlando, Florida, USA, 1999, pp. 77–84.

[30] M. Bartsch and G. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. of IEEE WASPAA*, New Paltz, NY, USA, 2001, pp. 15–18.

[31] G. Peeters, "Modeles et modelisation du signal sonore adaptes a ses caracteristiques locales," PHD Thesis, Universite Paris VI, 2001.

[32] B. Doval and X. Rodet, "Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and hmms," in *Proc. of IEEE ICASSP*, vol. 1, Minneapolis, USA, 1993, pp. 221–224.

[33] A. deCheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[34] R. Maher and J. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *J. Acoust. Soc. Am.*, vol. 95, no. 4, pp. 2254–2263, 1994.

[35] C. Yeh, A. Roebel, and X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," in *Proc. of IEEE ICASSP*, vol. 3, Philadelphia, PA, USA, 2005, pp. 225–228.

[36] P. Herrera, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer-Verlag, 2006.

[37] MIREX, "Music information retrieval evaluation exchange," 2005, 2006, 2007,2008.

[38] G. Peeters and X. Rodet, "Hierachical gaussian tree with inertia ratio maximization for the classification of large musical instrument database," in *Proc. of DAFX*, London, UK, 2003, pp. 318–323.

[39] M. Hall, "Feature selection for discrete and numeric class machine learning," Tech. Rep., 1999.

[40] G. Box and D. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society*, pp. 211–252, 1964.

[41] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[42] E. Frank and I. Witten, "Generating accurate rule sets without global optimization," in *Proc. of ICML (Int. Conf. on on Machine Learning)*, 1998, pp. 144–151.

[43] E. Frank, L. Trigg, M. Hall, and R. Kirkby, "Weka: Waikato environment for knowledge analysis," 1999-2000.

[44] L. Rabiner, "A tutorial on hidden markov model and selected applications in speech," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[45] O. Izmirli, "Using a spectral flatness based feature for audio segmentation and retrieval," in *Proc. of ISMIR*, Pymouth, Massachusetts, USA, 2000.

[46] MPEG-7, "Information technology - multimedia content description interface - part 4: Audio," 2002.

[47] D. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast," in *Proc. of ICME (IEEE Int. Conf. on Multimedia and Expo)*, Lausanne Switzerland, 2002.

[48] G. Wakefield, "Mathematical representation of joint time-chroma distributions," in *Proc. of SPIE conference on Advanced Signal Processing Algorithms, Architecture and Implementations*, Denver, Colorado, USA, 1999, pp. 637–645.

[49] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," in *Proc. of ICMC*, Bejing, China, 1999, pp. 464–467.

[50] S. McAdams, S. Windsberg, S. Donnadieu, G. DeSoete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: common dimensions specificities and latent subject classes," *Psychological research*, vol. 58, pp. 177–192, 1995.

Geoffroy Peeters Geoffroy Peeters received his Ph.D. degree in computer science from the Universite Paris VI, France, in 2001. During his Ph.D., he developed new signal processing algorithms for speech and audio processing. Since 1999, he works at IRCAM (Institute of Research and Coordination in Acoustic and Music) in Paris, France. His current research interests are in signal processing and pattern matching applied to audio and music indexing. He has developed new algorithms for timbre description, sound classification, audio identification, rhythm description, automatic music structure discovery, and audio summary. He owns several patents in these fields. He has also coordinated indexing research activities for the Cuidad, Cuidado, and Semantic HIFI European projects and is currently leading the music indexing activities in the Quaero Oseo project. He is one of the co-authors of the ISO MPEG-7 audio standard.

Emmanuel Deruty Emmanuel Deruty graduated from CNSMDP in 2000 (Tonmeister course). He worked at IRCAM from 2000 to 2003 with Louis Dandrel in the Sound Design service. He worked as freelance sound designer and music composer in Paris, San Francisco, NYC and London from 2003 to 2007, then joined IRCAM back in 2008 for the music part of the Quaero Oseo project.