euronoise

Automatic morphological description of sounds

Geoffroy Peeters and Emmanuel Deruty

Ircam, 1, pl. Igor Stravinsky, 75004 Paris, France
peeters@ircam.fr

Morphological description of sound has been proposed by Pierre Schaeffer. Part of this description consists in describing a sound by identifying the temporal evolution of its acoustical properties to a set of descriptors. This kind of description is especially useful for indexing sounds with unknown cause such as SoundFX. The present work deals with the automatic estimation of these morphological descriptions from audio signal analysis. In this work, three morphological descriptions are considered: - dynamic profiles (ascending, descending, ascending/descending, stable, impulsive), - grain/ iteration profiles, - melodic profiles (up, down, fixed, up/ down, down/ up). For each case we present the most appropriate audio features and mapping algorithm used to automatically estimate these profiles. We demonstrate the use of these descriptions for automatic indexing and search-by-similarity.

1 Introduction

Sound description has been the subject of many researches over the last decades. Most of the researches in this field focus on the recognition of the sound source (the cause that has produced the recorded sound). For example [10] [5] [15], [6] propose systems for the automatic recognition of musical instruments (the cause of the sound), [8] for percussive sounds. Other systems focus on describing sound using the most perceptually significant characteristics (based on experimental results). For example [20] [11] [17] [1] [3] 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 final results are given to him. This is because it is difficult to share a common language for sound description [7] outside the usual source/ causal description. Therefore, a problem arises when dealing with abstract sounds, SoundFXs, unnatural or synthetic sounds for which the source/ cause is usually unknown or unrecognizable. Another approach must be used for these sounds.

In this paper we propose a system for generic sound description based on Pierre Schaeffer proposals. In "Traite des objets musicaux" [19] (later reviewed by [2]), 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 we define 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 (polyphony).

1.1 Morphological sound description

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 [18] into

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

1.2 Ecrins' sound description

Following Schaeffer works, there has been much discussion concerning the adequacy or not of the proposed 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 Schaeffer criteria [13] [9].

In the Ecrins project (Ircam, GRM, Digigram)[12], a set of criteria based on Schaeffer work has been established for the development of an online sound search-engine. The search-engine must use sound description coming from automatic sound indexing. In this project, the morphological criteria (called morphological sphere) are divided into two descriptors sets: main and complementary [4].

The **main descriptors** are: - the duration, - the dynamic profile (flat, increasing or decreasing), - the melodic profile (flat, 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).

Icons representing the main-descriptors have been integrated in a Flash-based interface. This interface allows the user to enter easily the description of new sounds or to create a query based on specific morphological criteria (see Fig. 1).

1.3 Paper content and organization

The present work deals with the automatic estimation of this morphological description from audio signal analysis. Among the proposed descriptions three morphological descriptions are considered:

- Dynamic profiles (ascending, descending, ascending/descending, stable, impulsive),
- Grain/ iteration profiles,
- Melodic profiles (up, down, fixed, up/ down, down/ up).

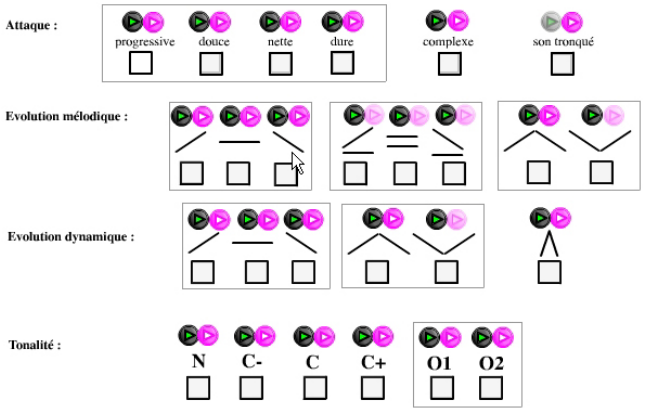

Figure 1: Flash interface for iconic representation of the main morphological sound descriptors.

For each case we present the most appropriate audio features and the mapping algorithm used to automatically estimate the profiles.

The paper is organized as follows. Part 2 present the concept of audio features and the inadequacy of the usual feature temporal models to represent morphological profiles. Parts 3.1, 3.2, 3.3 present the algorithms developed to estimate the three considered morphological profiles. We finally discuss the results in part 4 and present further works.

2 Sound description

2.1 Audio features

An audio feature (sound descriptor) is a numerical value which describes a specific property of an audio signal. Most of the time, audio features are extracted by applying signal processing algorithms (FFT, Wavelet...) to an audio signal. Depending on the audio-content (musical instrument sound, percussion, SoundFX, speech, music...) and on the application (indexing, search-by-similarity) numerous audio features have been proposed: spectral centroid, log-attack-time, Mel frequency cepstral coefficients... A list of the most commonly used audio features can be found in [16].

2.2 Modelling time

Audio features are usually extracted on a frame-basis: a value is extracted every 20ms. These features are called "instantaneous". A sound is then represented by the succession of its instantaneous features. 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 delta-features or acceleration features. The features can also be summed up using their statistical moments over larger period of time (by computing the mean and variance of instantaneous features over a 500ms sliding-window). These features are often called "texture window". The notion of "succession" can also be represented using time-dependent statistical models such as hidden Markov models.

Usual audio indexing problems are solved by computing instantaneous features, modelling their large-scale statistical moments and then applying pattern matching algorithms (GMM, HMM...). This approach is known as the "bag-of-frame" approach. However, when applied to the problem of morphological profiles description, this approach leads to poor results. This is because usual temporal modelling methods do not allow matching the shape of the considered morphological profiles.

In the present work, instead of using generic audio features and use them to train complex statistical models, we develop specific (complex) audio features which allow distinguishing classes with simple statistical models (linear separability). 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 a test-set for each morphological profile.

3 Morphological sound description

3.1 Dynamic profiles

The input sound of the system is supposed to be unitary (i.e. cannot be segmented further). The sound is also supposed to belong to one of the five considered dynamic profiles: - ascending, - descending, - ascending/ descending, - stable, - impulsive.

3.1.1 Feature extraction

On feature design: Usual sound description systems works a-posteriori: they try a-posteriori to map extracted audio features to a sound class definition. We work the opposite way using an a-priori approach: we develop audio features corresponding directly to the considered classes (the five profiles).

Loudness: Since the dynamic profiles are related to the perception of loudness, we first extract the instantaneous AudioLoudness $l(t)$ from the signal. We then use this time function to estimate the various dynamic morphological profiles.

Slope estimation: The profiles "ascending", "descending", "ascending/ descending" are described by estimating the slope 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 $l(t)$ (low-pass filtering). We then compute two slopes: one before and one after t_M.

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.

Time normalization: The dynamic profiles must be independent of the total duration of the sound (a sound can increase over 1s or over 60s, it is still an "increasing" sound). For this, all the computations are done on a normalized time axis ranging from 0 to 1.

B-spline approximation: In order to get the slope corresponding to the dynamic profiles we want to approximate $l(t)$ by two first-order polynomial 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 $(t_f, l(t_f))$, $(t_M, l(t_M))$ and $(t_l, l(t_l))$. 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 guarantee to connect at t_M.

Effective duration: The two-slopes model allows to represent the "ascending", "descending", "ascending/

descending" profiles as well as the "stable" profile (in this case the two slopes are equal and small). The distinction between "impulsive" profile and the other ones is done by computing the TemporalEffectiveDuration of the signal. The TemporalEffectiveDuration is defined as the time $l(t)$ is above a given threshold (40% in our case), normalized by the total duration. The various stages of the extraction are summed up here (see also Fig. 2):

1. Extraction of the instantaneous AudioLoudness $l(t)$,
2. Apply a low-pass filter to $l(t)$,
3. Apply a threshold to $l(t)$ equal to 10% of $l(t_M)$,
4. Locate the maximum value of $l(t)$,
5. Express $l(t)$ in log-scale,
6. Model $l(t)$ using a second-order B-spline,
7. Convert the B-spline to PP-spline.

In Fig. 3 we illustrate the extraction process on a real signal belonging to the "ascendant" dynamic profile.

From the spline approximation we compute the following set of features (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, - ED: TemporalEffectiveDuration of the whole signal.

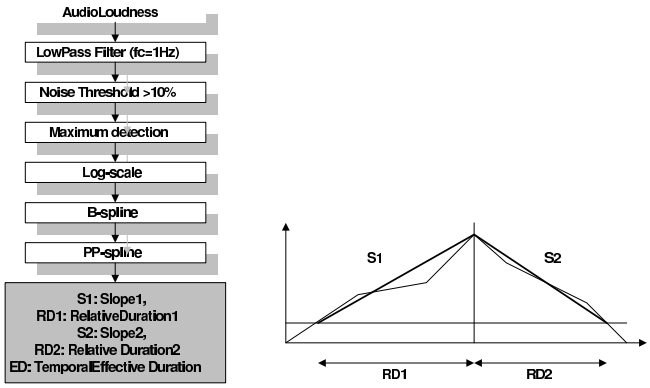

Figure 2: Audio features extraction algorithm for dynamic profiles estimation.

3.1.2 Evaluation

We have evaluated the proposed extraction method on a test-set of 187 audio files (26 ascending, 68 descending, 24 ascending/ descending, 37 stable, 32 impulsive). The sounds are part of the "Sound Ideas 6000" collection and have been selected by one of the author only based on their perceptual characteristics.

In Fig. 4, we represent the 187 sounds in the feature spaces x=S1/ y=RD1 (top part), x=S2/ y=RD2 (bottom part). The dynamic profiles classes are represented as: ascending (green), descending (red), ascending/ descending (black), stable (magenta), impulsive (blue). Fig. 4 clearly shows a separation of the four first classes. In this figure, the impulsive (blue) and ascending/ descending (black) classes are mixed; this is because the TemporalEffectiveDuration is not represented here.

We have also performed a classification test using the PART algorithm. The PART algorithm provides

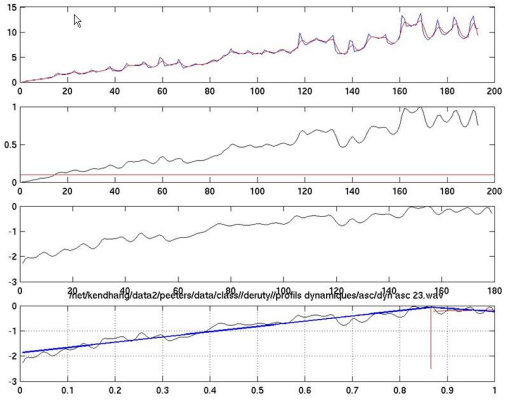

Figure 3: Estimation of dynamic profile parameters: a) loudness (black) and smoothed loudness over time (red), b) 10% threshold apply to smoothed loudness, c) smoothed loudness in log-scale, d) Maximum value (vertical red bar) and B-spline modeling.

the set of rules (indicated into Tab. 1) that must be used to perform automatic classification.

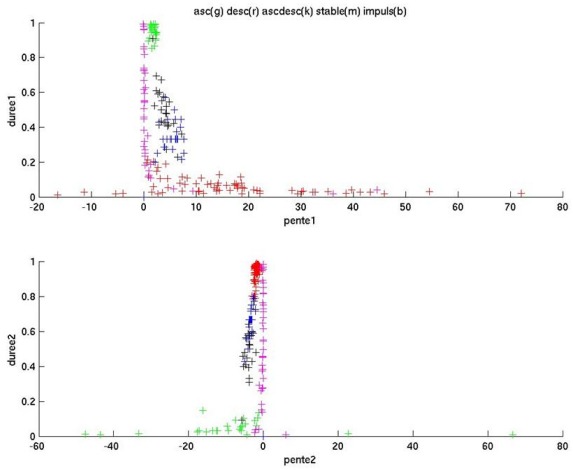

Figure 4: Representation of the dynamic profile test-set into the the feature spaces x=S1/ y=RD1 (top part), x=S2/ y=RD2 (bottom part).

3.2 Grain/ iteration profiles

Six grain/ iteration profiles have been considered: - IT: periodic, - IHV: periodic with variable intensity, - IHT: periodic with variable timbre, - IV: variable iterative, - IPH: periodic with non-periodic elements inserted, - IPV: periodic layer + non-periodic layer.

The profiles are illustrated by a set of 188 sounds coming from the "Sound Ideas 6000" collection (67 IT, 4 IHV, 41 IHT, 10 IV, 2 IPH, 4 IPV + 22 IV-IHT, 1 IV-IPH, 5 IV-IPH-IHT). The sounds have been selected by one of the author only based on their perceptual characteristics.

3.2.1 Feature extraction

Grain/ Iteration profiles: 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

Descending	ED > 0.28 AND Dur1 <= 0.214286 AND Slope2 <= -0.584657
Stable	ED > 0.28 AND Slope2 > -1.287156 AND Slope2 <= 6.022541
Impulsif	ed <= 0.28: (32.0)
Ascending	Dur1 > 0.73991 AND Slope1 > 0.403746
Ascending/Descending	Slope1 > 0.972048

Table 1: Rules for automatic classification into dynamic profiles.

complicates the automatic detection of the repetition. Moreover several repetition cycles can occur at the same time for the same parameters (given 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 MFCC. Another complexity comes from the variation of the period of repetition over the sound duration or from disturbance from other perceived parameters.

Among the wide range of possible descriptions for iterative sounds we selected the three following predominant characteristics which are connected to the six considered profiles:

- The **amount** of repetition: allows distinguishing between iterative sounds and non-iterative sounds
- The **period** of the cycle: allows distinguishing between "grains" (short period) and "repetitions" (long period)
- The **characteristics** of the repeated element: allows distinguishing between percussive elements and smooth elements

The algorithm works in three stages: 1) Estimation of the amount of periodicity of the sounds and estimation of the period of the cycle, 2) localization of one of the repeated element, 3) characterization of the repeated element.

Estimation of the periodicity for dynamic elements: The AudioPower descriptor is computed with a small hop size (5ms). This small hop size allows the description of fast repetition (such as "grains") or slow repetition. The amplitude spectrum of this energy function is computed and the maximum peak of the spectrum in the range [0.1, 20] Hz is estimated. The period corresponding to this frequency is considered as the period of the cycle. The amount of periodicity is given by the value of the normalized auto-correlation function at the period of the cycle position.

Estimation of the periodicity for timbre/pitch elements: The AudioMFCC descriptor is computed. The corresponding similarity matrix is computed and transform to the lag-domain (lag-matrix). An AudioSimilarity function is computed as the normalized sum over the time axis of the lag-matrix (sum over the column normalized by the number of element in each column of the matrix). This AudioSimilarity function expresses the amount of similarity of the sounds for specific lags. The amplitude spectrum of the AudioSimilarity function is computed and the maximum peak of the spectrum in the range [0.1, 20] Hz is estimated. The frequency of this peak is considered as the period of the cycle. The amount of periodicity is given by the value of the normalized auto-correlation function at the period of the cycle position.

Localization of one of the repeated element: The localization of the repeated elements is done by a method developed for PSOLA pitch-periods localisation [14]. Given the period of the cycle T, we define a vector of cycle instants $\underline{T_\tau(t)} = \sum_k \delta(t - \tau - kT)$ ($\underline{T}$ is a comb-filter starting at time τ with periodicity T). The local-maxima of the AudioPower (or of the AudioSimilarity) around the values of $\underline{T}$ are detected. We compute the sum of the AudioPower at these positions. The process is repeated for various τ values. The value τ leading to the maximum sum defines the vector which gives the best time locations for a segmentation into repeated elements.

Characterization of the repeated element: One of the sound-elements is then used for characterizing the signal in terms of perceptual audio features. The following audio features are extracted for this element (see [16] for details on the audio features): - TemporalIncrease, - TemporalDecrease, - TemporalEffectiveDuration, - AudioSpectrumCentroid, - AudioSpectrumSpread.

The flowchart of the extraction process is illustrated into Fig. 5.

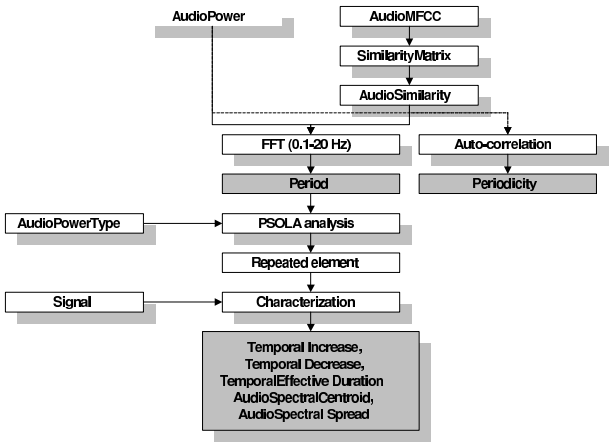

Figure 5: Audio features extraction algorithm for iterative profiles estimation.

3.2.2 Evaluation

Since the proposed description (amount of repetition, period of the cycle and characteristics of the repeated element) does not match directly the six proposed grain/ iteration profiles of the test-set, it hasn't be possible to perform a classification evaluation for the grain/ iteration description.

The grain/ iteration description has however been used in a prototype search-by-similarity application. In this application the user can select a sound and ask for sounds with similar iteration speed and/or audio characteristic of the sound-elements. Each of the criteria used for the search-by-similarity can be weighted between 0 and 1 in order to de-emphasize or emphasize a specific criterion.

3.3 Melodic profiles

Five melodic profiles have been considered: - up, - down, - fixed, - up/ down, - down/ up.

The profiles are illustrated by a set of 188 sounds coming from the "Sound Ideas 6000" collection (71 up, 56 down, 32 fixed, 23 up/ down, 6 down/ up). The sounds have been selected by one of the author only based on their perceptual characteristics.

Despite the shared perception of the melodic profiles, it has not been possible to create a system for the automatic extraction of them. This comes from several reasons:

- Pitch/ Spectrum: the perception of melodic profiles comes either from a modification of the pitch or from a modification of the spectral envelope (spectral centroid, resonant frequency). The current profiles should therefore be further divided into sub-profiles according to these characteristics.
- Time extend: some profiles judged as ascending melody are in fact decreasing (in spectral content) over a long period of time and only increasing (in pitch) over a small period of time.
- Non-unitary melodic profiles: other profiles judged as ascending melody are in fact quick repetition of descending-note-arpeggio (such as in harp sounds) with increasing root-note.

In conclusion, in order to be able to apply automatic extraction algorithms for the melodic profiles estimation, further refinements are first needed in the specification of these profiles.

4 Conclusion

In this paper we have presented audio features and mapping algorithms for the automatic estimation of three morphological profiles derived from Schaeffer proposed description of sound.

The dynamic profiles estimation is achieved using temporal loudness estimation which is then approximated using B-splines. The extracted features allow a good match with a provided test-set.

The grain/ iterative profiles are described by the amount of periodicity of the signal, the periodicity itself and an acoustical description of the repeated elements. Spectral analysis of the AudioPower function or of the AudioSimilarity derived from the AudioMFCC Lag-matrix are proposed to measure the periodicity. A PSOLA algorithm is then apply to locate the repeated elements which are then described by a set of perceptual audio features. The extracted features were used in a query-by-example application with positive user-feedbacks.

We finally discussed the case of the melodic profiles and highlighted the problems with the current categorization of the profiles which does not allow their automatic estimation. Further work will therefore concentrate on that.

The remaining descriptions of sound objects presented in Part 1.2 (attack, pitch and spectral distribution) were not discussed in this paper since they do not involve modelling time. These descriptions can be obtained using the audio features described in [16] and were discussed in previous works such as [18].

5 Acknowledgments

Part of this work was conducted in the context of the Ecoutes French Project, CUIDADO I.S.T. European Project and Sample Orchestrator ANR French Project.

References

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

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

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

[4] E. Deruty. Ecrins report: Descripteurs morphologiques / sons essentiels. Technical report, Ircam, 2001.

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

[6] S. Essid. *Classification automatique des signaux audiofréquences: reconnaissance des instruments de musique*. PhD thesis, 2005.

[7] A. Faure. *Des sons aux mots : Comment parle-t-on du timbre musical ?* Phd thesis, Ecoles des Hautes Etudes en Sciences Sociales, 2000.

[8] P. Herrera, A. Dehamel, and F. Gouyon. Automatic labeling of unpitched percussion sounds. In *Proc. of AES 114th Convention*, Amsterdam, 2003.

[9] R. Leblanc. *Elaboration d'un systeme de classification pour sons figuratifs non instrumentaux*. Dess thesis, Universite Pierre et Marie Curie, Paris 6, 2000.

[10] K. Martin. *Sound source recognition: a theory and computational model*. Phd thesis, MIT, 1999.

[11] 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, 1998.

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

[13] C. Olivier. *La recherche intelligente de sons*. Master thesis, Univ. de Provence, France, 2006.

[14] G. Peeters. *Modeles et modelisation du signal sonore adaptes a ses caracteristiques locales*. Phd thesis, Universite Paris VI, 2001.

[15] 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. Peeters03b.

[16] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado i.s.t. report, IRCAM, 2004.

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

[18] J. Ricard and P. Herrera. Morphological sound description: Computational model and usability evaluation. In *Proceedings of AES 116th Convention*, Berlin, Germany, 2004.

[19] P. Schaeffer. *Traité des objets musicaux*. Seuil, Paris, 1966.

[20] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification search and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.