Semiotics of Sounds Evoking Motions: Categorization and Acoustic Features

Adrien Merer[1], Sølvi Ystad[1], Richard Kronland-Martinet[1], and Mitsuko Aramaki[2,3]

[1]CNRS - Laboratoire de Mécanique et d'Acoustique,
31 ch. Joseph Aiguier, Marseille, France
[2]CNRS - Institut de Neurosciences Cognitives de la Méditerranée,
31 ch. Joseph Aiguier, Marseille, France
[3]Université Aix-Marseille
38 bd. Charles Livon, Marseille, France
{merer,ystad,kronland,aramaki}@lma.cnrs-mrs.fr

Abstract. The current study is part of a larger project aiming at offering intuitive mappings of control parameters piloting synthesis models by semantic descriptions of sounds, i.e. simple verbal labels related to various feelings, emotions, gestures or motions. Hence, this work is directly related to the general problem of semiotics of sounds. We here put a special interest in sounds evoking different perceived motions.
In this paper, the experimental design of the listening tests is described and the results obtained from behavioural data are discussed. Then a set of signal descriptors is compared to categories using feature selection methods. A special interest is given to applications for sound synthesis.

Key words: sound semiotics, motion, categorization, sound synthesis

1 Introduction

In the sound design context, synthesizing sounds from verbal labels related to various sensations, emotions, gestures or motions is still an open problem. Also in a musical context, composers want to create or transform sounds by acting on parameters that are relevant from a perceptual point of view. Indeed, physical synthesis models often necessitate the manipulation of hundreds of parameters. Consequently the construction of "good" sounds is almost impossible if no mapping strategy is used. In addition, certain signal models like FM synthesis are hard to control even after a learning process since the relation between timbre and synthesis parameters is non-linear. Most approaches, consist in first building a synthesis model and then addressing the mapping between synthesis parameters and control parameters. Indeed, our approach, the so-called "semiotic"[1] approach, consists in building the synthesis model directly from the control parameters which are relevant from a perceptive/cognitive point of view.

[1] the study of signs and symbols, what they mean and how they are used, *Cambridge Advanced Learner's Dictionary*

This approach leads to the more general issue: understanding how listeners assign meanings to sounds and in particular, determining acoustic features that convey information.
Semiotics has been studied in several contexts such as music information retrieval [22], perception of impact sounds [1], noise annoyance [14], sound design [13], perception/cognition of romantic music [2]. In particular, in the context of product quality evaluation U. Jekosch [13] addressed a theoretical framework based on a general theory of signs and in accordance with Gestalt perception. She also explains that this approach is particularly relevant in an industrial context.

In this study, a general methodology based on 3 steps is proposed:

- Determination of sound categories;
- Determination of invariants representative of these sound categories;
- Control of synthesis processes based on these invariants (sonification).

Many aspects of sound are concerned by semiotics. Indeed, listening to the same sound, different listeners might focus on different information carried by the sound. Conversely, some information can be gathered by only a few listeners. For example listening to a voice through the telephone, you might detect different moods if you are familiar with the speaker or not. Hence, the information conveyed by sound studied through a semiotic approach should be as independent of listeners' "history" as possible. This leads to a consideration of only basic properties of sound sources (*e.g.* size, material, displacement...) experienced in everyday listening [6].
As a first attempt to identify signal parameters linked to a sound source property, we have here focused on the evocation of motion. Motion is a primordial aspect of the appreciation of music. Indeed, in [5], authors studied the association between musical parameters and images of motion, and identified important links between gesture and various parameters such as pitch, loudness and rhythm.

Following the general methodology presented above, the first step consisted in determining categories of sounds evoking motions by listening tests. For this purpose, sounds from data banks made by electroacoustic music composers were collected. Among the large number of samples, we chose sounds which sources cannot be identified, but which, however, convey a signification. This made it easier for the subjects to focus on the evocations induced by the sounds,without being influenced by the identification of sound sources. Indeed in [9], Guastavino observes that in the case of environmental sounds, listeners process sounds as semantic labels. The author also indicates that for "abstracted stimuli", obtained categories might be more correlated with acoustic features. Conversely, Schaeffer [21] (and others) assumed that when we listen to a sound, we automatically try to link this experience with a similar one, stored in our memory. In the case of electroacoustic music, we can for instance predict that listeners will make comparisons with audio effects used in science fiction movies. This comment brings us to consider the problem of context. Indeed according to the ecological approach of perception [7], [6] everyday listening is usually related to complex events. Besides, we must consider that sound perception is multi-modal. Listening tasks through headphones or loudspeakers (corresponding to listening conditions in laboratory) have been studied by Schaeffer [21] through what he calls the ”acousmatic” approach. Schaeffer explains that this approach permits to separate auditory and visual information and to make us aware of the fact that the listening changes over time when we repeatedly listen to a sound. In our daily-life a lot of information comes from loudspeakers in radio, TV, computers, alarm systems etc. These considerations lead us to the choice of stimuli from electro-acoustic music composer since they are complex and can refer to various sound source properties and are well adapted for listening tasks through headphones.

Synthesized sounds were also included in the sound material to integrate some assumptions related to the physics of moving sound sources. In practice, the following physical phenomena were simulated: Doppler effect (known to give the sensation of a passing source), air absorption (known to be important for the perceived distance of a source), reverb (known to be important for the perceived distance or for the sensation of room acoustics) and raise/decay of sound pressure level. We tested if sound transformations corresponding to each of these physical phenomena simulated independently can evoke specific motions.

To define categories from the collected set of sounds, we conducted 2 categorization tasks where participants were asked to group sounds as function of the evoked motions (or displacements). In the first experiment, participants were allowed to make as many groups as they wanted, whereas, in the second experiment, they had to group sounds in predefined categories, each of them being represented by a prototypical sound obtained from the results of the first experiment. This approach makes it possible to avoid verbal labels [2]. Free categorization has many advantages (compared to dissemblance tests for example) in the sense that a lot of stimuli can be tested. It gives simultaneously access to categories (with verbal descriptions) and corresponding sounds. In addition, no hypothesis about the existence of continuous perceptual dimensions is needed. Furthermore, we assumed that in the second task, the high variability of the results obtained in the first task will be reduced.

The categories of movements obtained from the behavioural data were further examined in order to identify signal features specific to each category. First, the analytical properties of each sound were calculated through several signal descriptors described in section 3. Then, statistical analysis lead to the most relevant descriptors (signal features) specific to each category of motion.

We finally discuss some perspectives concerning the control of these descriptors (last step of our methodology).

2 Determination of Sound Categories

2.1 Stimuli

Recorded Sounds We preliminary collected about one thousand samples from personal data banks belonging to electroacoustic composers of the Music Conservatory of Marseille, with their agreement. Sounds were all monophonic with 16-bit 48kHz sampling rate. These samples are essentially dedicated for musical compositions and are generally used as or after some audio effect transformations. Among these samples, a selection of 62 sounds was effectuated with respect to different criteria. First, according to the acousmatic listening context, we avoided caricatured sounds (like sounds used for cartoons) and sounds for which the sources were easily identifiable. Second, we restricted our selection to sounds that present a simple morphology (single event) and that last no longer than 4 seconds. We also cared that sounds should not be dramatically cut from a longer sample. This point is of importance since it can influence the categorization task if used as a strategy of comparison between sounds. Finally, according to analysis constraints, we aimed at constituting the most heterogeneous sound panel with respect to timbre, duration and level.

Synthesized Sounds Hypothesis about acoustic information related to a moving sound source are tested by including additional sounds obtained by transformation of 6 original recorded samples different from the 62 sounds previously selected. The original samples were first modified to freeze the evolution of signal parameters by using a phase vocoder freezing technique [19]. Then, we applied sound transformations corresponding to the following physical phenomena: air absorption, raise/decay of sound pressure level, reverb and Doppler effect.
Air absorption is simulated by a first order low pass filter with varying cut-off frequency (from 13-kHz to 30-Hz). The raise/decay phenomenon is simulated by a geometric $1/r$ evolution of the sound pressure level, where r is the distance between the source and the listener. The reverb effect is effectuated by an Olaf Matthes freeverb MSP object (freeverb is a Schroeder / Moorer reverb model) without damping, max room size and varying reverb rate. Finally, the Doppler effect is reproduced with a delay line. For a monochromatic delayed sound source $s(t-D_t) = e^{i\omega_s(t-D_t)}$ with a time varying delay time D_t, the instantaneous frequency ω and the frequency measured at the listener's location (Doppler shift) ω_D are given by:

$$\omega = \omega_s(1 - \frac{dD_t}{dt}) \quad ; \quad \omega_D = \omega_s(\frac{1+\frac{v_{ls}}{c}}{1-\frac{v_{sl}}{c}}) \tag{1}$$

where v_{sl} and v_{ls} are the relative velocities between the source and the listener. Therefore, for a static listener ($v_{ls} = 0$) and assuming that $v_{sl} << c$, the delay time is given by: $\frac{dD_t}{dt} = -\frac{v_{sl}}{c}$. In practice, 4 sounds were constructed to simulate these 4 physical phenomena independently. In particular, reverb effect and air absorption are computed for a source approaching the listener with constant speed. The sound pressure level raise/decay and Doppler frequency shift are computed for a linear uniform movement of a sound source going past a fixed listener from -50 to 50 meters in 6 seconds. Two sounds were also constructed (with independent time dilation/compression and level variation) to simulate a rotating sound source around a listener located close to a 9 meter radius loop with an angular velocity of 18 tr/min.

2.2 Test 1: Free Classification Task

Twenty-six students (9 females, 17 males) working on the CNRS campus in Marseille participated in the experiment. They were between 19 and 30 years old (average 23,5), 19 had music experience (2 also had electroacoustic music experience).

Procedure

The listening tests were conducted in an audiometric cabin. Participants were placed in front of an imac computer screen and listened to **monophonic** sounds through a Stax 3R202 headphone set under binaural conditions with a SRM310 preamplifier (we used the internal sound card).

The 68 sound samples represented by square symbols, were initially positioned randomly on the screen. The task consisted in grouping together sounds evoking the same motion. Participants could listen to sounds and move them with the mouse as often as they wanted. We did not impose constraints about the number of categories to make and we insisted on the fact they should avoid identifying the nature of the sources that produced the sounds.
A training phase was effectuated for the participants to adopt the ecologic listening and focus their attention on the impression of motion evoked by sounds. This preliminary test allowed us to check if the participants were able or not to make abstraction from the sound source identification and if they well understood the instructions.
At the end of the task, participants were asked to describe (by sentences or a few words) the type of motion they associated with each group they formed on the screen. They finally wrote their global impression of the test (whether the task was hard or boring, the choice of sound material, etc ...).

Results

The test lasted from 21 to more than 60 min across participants. Except for one, all of them were satisfactory about the groups they made. As expected, we observed a high inter-subject variability in the number of categories. Indeed, participants formed in average 8.8 groups (standard deviation: 3.9), but the number varied from 3 to 21 groups across participants. We noted that six participants formed groups composed of only one or two sounds. One subject gave up the test, since no categories had been formed after 45 minutes and the screen was similar to its initial state.

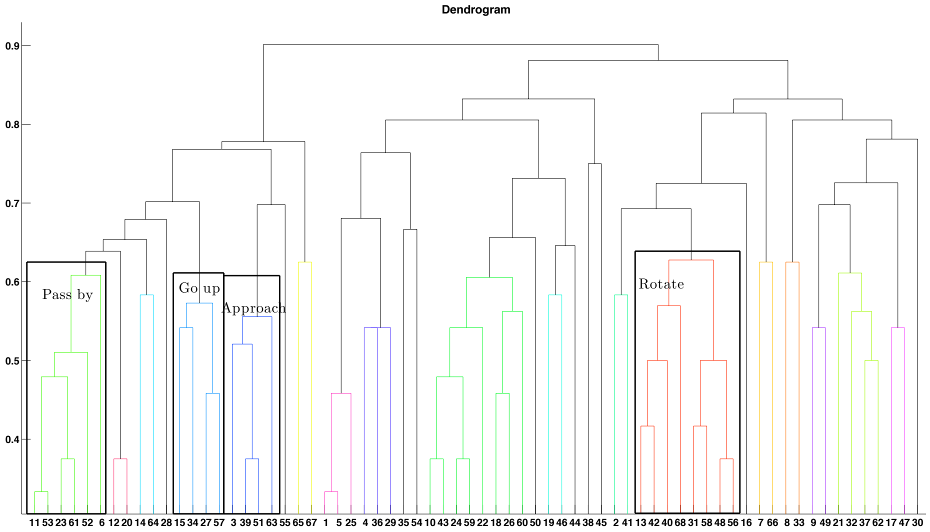

Fig. 1. Each sound (labelled with a number from 1 to 68) is linked to another according to their similarity. Each group of two sounds is linked to the closest until all sounds are linked together

Definition of the Most Representative Motion Categories

To highlight the most representative categories of evoked motions reflecting the participants judgement, we matched the results obtained by different (semantic and statistical) analyses of the behavioural data.

The first analysis was effectuated on the responses to the questionnaire filled by participants at the end of the listening test. In particular, words used by participants to describe the groups they formed, were compared across participants. As assumed, they used different words to describe a same evoked motion. In practice, groups which were described with similar words (synonyms) are considered together and we retained the most relevant label (following our own judgement) for each group.

We excluded complicated expressions or metaphors, which would necessitate detailed linguistic analysis. Hence, we identified six categories corresponding to the following motions: “rotate”, “fall down”, “approach”, “pass by”, “go away” and “go up”. Respectively 69%, 54%, 46%, 46%, 46% and 34% of the participants proposed these categories. We also extracted sounds corresponding to those categories according to the number of time they have been cited. Many sounds belong to different categories at a time since four of the six categories have been made by less than 50% of the subjects. Despite this, at least one sound appears more than 70% of the time for each category.

The second analysis was made on the participant’ classifications. The results were represented by a 68 × 68 similarity matrix where each cell indicates the percentage of participants that did group together the two sounds. The hier-

archical clustering analysis was conducted on the dissimilarity matrix obtained by subtracting the similarity matrix from 1. This method consists in linking together pairs of sounds with respect to their similarity, then linking these pairs with other pairs until all elements are grouped together. The obtained dendrogram is represented on Figure 1.
Finally, the results of the 2 analyses were set side by side to determine our most representative categories of motion. For that purpose, the similarity matrix was resorted according to the dendrogram (cf. fig2). It allows highlighting five groups that match five of the groups defined by semantic analysis of subjects words. For example the first six elements of the dissimilarity matrix contain the six sounds which have been cited by more than 50% of the subjects who made the category called "pass by".
Consequently, these five groups which were found both in semantic and cluster analysis, were defined as the main motion categories for our study. These groups were reported on the dendrogram of Figure 1.

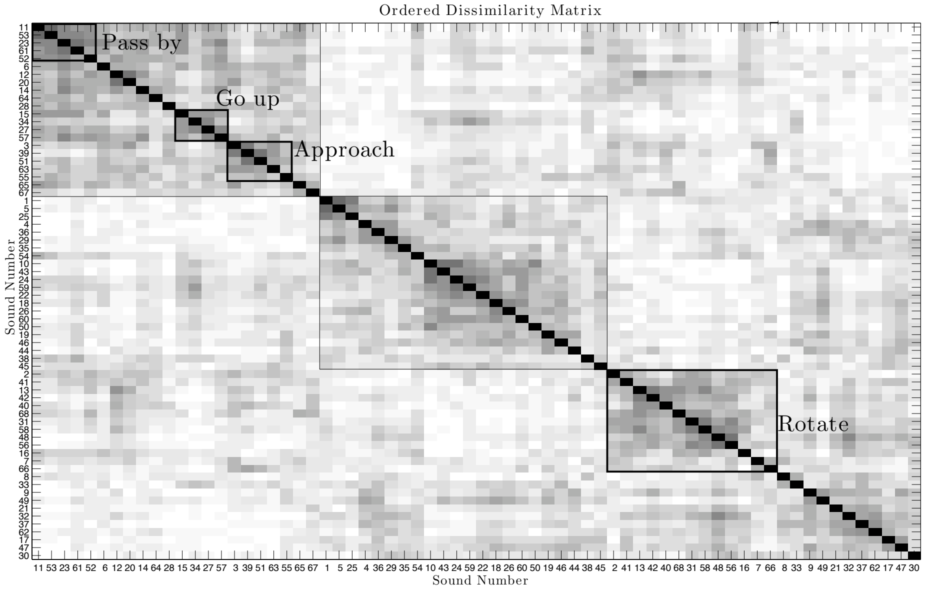

Fig. 2. Similarity matrix resorted according to the dendrogram. The grey scale corresponds to percent of time that two sounds are grouped together. Black: 100% White: 0%

Finally, for each group, we determined a "prototypical" sound representative of the category as sounds which have been cited at least by 70% of the subjects and not cited in another category. The five categories of motions are further used in test 2 as predefined categories.

2.3 Test 2: Restricted Classification Task

Sixteen subjects (6 females, 10 males) participated in this experiment and all of them participated in the first one (within a break of two weeks between the tests). Test 2 was conducted in the same experimental conditions than test 1. The task consisted in classifying the same stimuli into predefined categories of motion. These predefined categories were deduced from the most representative ones obtained from test 1. On the graphical interface, the top half of the computer screen was split in five boxes corresponding to these predefined categories. Sounds to be categorized were randomly located in the bottom half of the screen. Instead of labelling the predefined categories with a word, we characterized each of them by the prototypical sound that was defined from the results of test 1. Participants placed sounds from the bottom of the screen into one of the boxes as function of evoked motions. They also were allowed to leave sounds that they judged ambiguous on the bottom of the screen.

Results We computed the percentage of time each sound was sorted in each category of motion. In each category, sounds were ordered as function of their occurrence frequency. Thus, we arbitrarily fixed a threshold value at 70% beyond which sounds are defined as typical for the category[2]. With such a threshold, no sounds are representative of the category “approach”, 2 are representative for “rise”, 5 for “fall down” and “pass by” and 9 for the category ”turn”. In a further step, this threshold value has to be adjusted according to the number of sounds needed for the determination of the invariants of each category.
Most participants left some sounds at the bottom of the screen, but 62% answered ”yes” to the question ”Was the number of categories sufficient?”. Only 2 sounds are sorted in no category more than 50% of the time.

2.4 Comparison Test 1/Test 2

Test 2 gives groups that are valid for all the participants opposite to the first test in which only two groups where valid for more than 50% of the subjects. 70% found that the second test was easier than the first one and the time to complete the task were considerably lower in the second test (average 19 min for the second 43 min for the first).
Differences between the subjects’ answers to the first and to the second test are 23% (average of difference for each subject). The consistency between subjects’ answers is not higher in test 2. This is most likely linked to the fact that the participants focused on different aspects of the sounds and therefore associated different motions to them. Hence, the same sound can evoke motions such as rotate, go away and rise at the same time. This shows that even if our sound selection was supposed to exclude such complex sounds, selection is indubitably subjective (i.e. depends on researcher’s choice). The second test did not give the opportunity to associate more than one motion to each sound.

[2] see `http://www.sensons.cnrs-mrs.fr/CMMR07_semiotique/` for sound examples

2.5 Physical Considerations

The two sounds simulating the Doppler effect and raise/decay phenomena for a linear movement were not categorized together. Indeed, the second was typical for the category ”pass by” whereas the first was not sorted in this category (same comment for rotating sound source simulation). Indeed, according to Lufti & al. [15], the most significant cues for the perception of displacement of moderate velocity (10m/s) are intensity and inter-aural time difference. For high velocity displacements, the most significant cue is related to the perception of frequency shift due to the Doppler effect. Hence, cues used to perceive a source displacement seem to differ as function of the variation range of the velocity. To go further, it is important to see that such transformations are not always efficient to give an impression of motion. There is another example in the category ”pass by”, where we point out a sound with increasing centroid. This variation is in opposition to low pass filtering due to air absorption (and also to pitch shift due to Doppler effects) for a going away sound source, but 72% of the subjects described this displacement to be approaching and then going away.

3 Determination of Relevant Signal Descriptors of Categories

The listening tests allowed us to determine the 5 most representative categories of motion (“approach”, “rise”, “fall down”, “pass by” and ”turn”) and a set of associated typical sounds. In this section, we aim at determining the acoustic descriptors that are relevant to characterize each sound category. The problem to be solved here is quite analogous to automatic music classification in the sense that we want to find descriptors that can explain classification done by listeners. The field of music information retrieval gives lots of different answers to this problem (for example [4], [12]).
Most researches within automatic classification describe the same generic problems:

- Select the descriptors among thousand of possible
- Relevant criteria for evaluation of descriptors
- Robustness of model (which can be validated on other sound databases)

Often, the criteria are based on minimization between predictions of the model and experimental values. For example, the database is often split in two parts, the first one is used to build the model and the second for the evaluation. See [22] for a detailed analysis and complete overview of such problems.

It is important to keep in mind that we want to conduct our study while conserving as far as possible a general view of all the problems related to the semiotics of sounds in the context of synthesis. Afterwards we will be able to optimize our methodology. Indeed, as a first step, we focus on the determination of most relevant descriptors for each category instead of building a predictive model. Hence, with the results of listening tests in mind we can assert to build a model for which the validity is correlated with the number of data (between 4 and 8 among 68 sounds for each category).

3.1 Signal Descriptors

As Pachet and Roy [20] discuss, there are two different ways of selecting features: “by hand” and systematic selection. “By hand” means arbitrary selection of features according to common sense and systematic means algorithmic selection with no a priori. Our approach consists in selecting some well-known descriptors without assumption and building some others that seem to be relevant, and finally find a criterion to select the most relevant.
In most studies about timbre, the authors have developed signal descriptors to fit their perceptual dimension. Such descriptors are generally specific to musical instrument sounds, that is to say for quasi-harmonic spectra (for example in [8]). In our case, an important part of the ”corpus” is composed of noisy and non-stationary sounds. Hence we cannot use traditional auditory models to take into account loudness and masking.
In this study, we focused on a dynamic problem since the evoked movement is linked to a temporal evolution of the sound. For this reason we extend our analysis to descriptors that would quantify these temporal behaviours.

To characterize our sounds from an acoustic point of view, we calculated some well-known descriptors (spectral centroid, spectral spread, spectral variation, energy envelope, temporal centroid and signal duration). Since most sounds contain stochastic contributions, the spectral descriptors are calculated from the power spectrum density (PSD). Those descriptors are calculated with a frame based method (Hanning windows of 2048 samples with 50% overlapping between two successive frames).

Since we aim at characterizing the evolutional aspect of the sounds and make these evolutions comparable across sounds, we reduce time dependent descriptors to scalars. Hence we compute average, standard deviation, monotonousness and variation rate which are described below.

Monotonousness is defined as

$$Mn = \frac{1}{N} \sum_{n=1}^{N} sign\big(va'(n)\big) \tag{2}$$

where $va'(n)$ is a derivative of a discrete variable $va(n)$ of length N. Hence $Mn \approx 1$ means an increasing curve (resp. decreasing for -1) and oscillating or horizontal if $Mn \approx 0$. Monotonousness describes both curve variation sign and curve flatness.

Variation rate is defined as :

$$Vr = \frac{1}{N-1} \sum_{n=1}^{N-1} abs\Big(sign\big(va'(n+1)\big) - sign\big(va'(n)\big) \Big) \tag{3}$$

Thus this is the zero crossing-rate of signal derivative, which is correlated with the second order moment. It makes it possible to characterize smoothness of the curve independently from global evolution (as opposed to monotonousness).

Spectral Centroid Spectral centroid is one of the most known and used descriptors. This measure of the gravity center of the spectrum is closely related to the brightness of a sound. We used the definition proposed by Grey and Gordon [8]:

$$Sc = \sum_{k=1}^{K} \frac{k c_k}{\lambda + \sum c_k}$$

where c_k are coefficients of discrete PSD computed on frequency k and λ is a regulation parameter.

Spectral Spread is a measure of the spread of a spectrum around its mean value and can be calculated through the second order moment of the spectral centroid. From the calculation of Sc, we compute the equivalent of the second order moment (definition from Peeters [18]):

$$Ss = \sqrt{\frac{\sum_k (k - Sc)^2 c_k}{\sum_k c_k}}$$

Spectral variation (or Spectral Flux) is a measure of the time evolution of the spectrum and is defined by Peeters [18]

$$Sv(n) = 1 - \frac{\sum_k c(n-1,k) \times a(n,k)}{\sqrt{\sum_k c(n-1,k)^2} \sqrt{\sum_k c(n,k)^2}} \tag{4}$$

where $c(n,k)$ is PSD of the signal computed at the n^{th} frame.

Amplitude Envelope The envelope $A(n)$ is calculated by first computing the Hilbert transform $\mathcal{H}$ of the temporal signal and then by applying to its modulus, a low pass filtering (second order Butterworth filter) with cut-off frequency

fc. This filter determines the time scale for the energy variation. Fluctuation strength [24] is defined for amplitude modulations under 20Hz, and we therefore use this value as cut-off frequency to get a measure of the variations in this domain.

Temporal Centroid is the energy envelope centroid from definition of [18] (but normalized by signal duration):

$$Tc = \frac{1}{N} \frac{\sum_{n=1}^{n=N} n \times A(n)}{\sum_{n=1}^{N} A(n)} \tag{5}$$

Characterization of Amplitude Modulation The physics of moving sources states that periodicity is a fundamental characteristic of rotating sources. Thus we added to our set, descriptors that characterize amplitude modulation taking into account some specificities.
Most methods that estimate periodicity are based on the autocorrelation, which is also the case for the algorithm presented here. We compute the autocorrelation of the temporal signal. In our case, we rather consider the autocorrelation on the amplitude envelope $A(n)$ as defined previously from which some ”static” components are removed. In practice, a linear or quadratic interpolation was estimated from the autocorrelation and subtracted from it. The calculation of the autocorrelation on the envelope (instead of directly on the temporal signal) allow focusing on the slowest periodicities contained in the signal. The periodicity is quantified by calculating the Fourier transform of this adjusted autocorrelation. Then we detect the most prominent peaks, taking into account the peaks’ width. Thus, wide peaks are excluded (with arbitrary threshold) since they do not correspond to a detectable modulation. Actually, the threshold corresponds to a width of 5Hz from 75% of the component energy (cf. figure3). We then extract the frequencies and amplitudes of the two highest peaks.
In order to characterize variations in the amplitude modulation, we also extract the number of peaks above the energy average in the considered bandwidth (0-20Hz). Indeed, for variations of the modulation frequency, the autocorrelation ”spectrum” contains more peaks and the domain containing peaks characterize the modulation boundary.

Characterization of Level Variation Another fundamental characteristic of signals given by the physics of moving sources is the variation of loudness within a time interval corresponding to the length of the sound. For example, this is particularly important as a distance cue for low speed moving source as shown in [15].

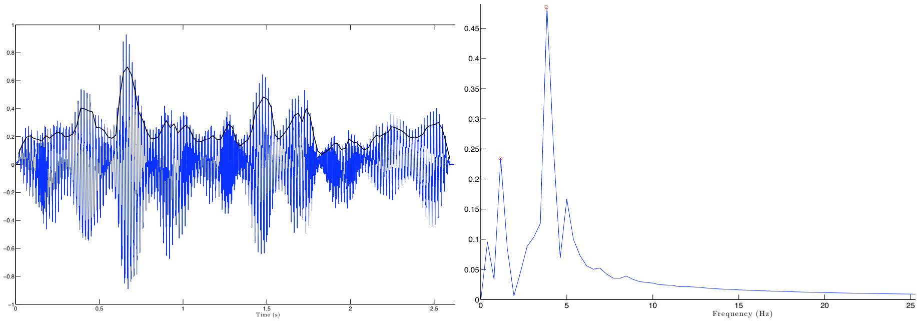

Fig. 3. Left: Estimation of the temporal envelope of a sound in the category "Rotate". Right: FFT of adjusted autocorrelation and extraction of peaks for the characterization of amplitude modulation of the temporal envelope. The selected peaks are marked by circles.

To characterize this level variation, we detect the maximum of the energy envelope and calculate a linear interpolation of the logarithm of the envelope from the beginning of the signal to its maximum value and further from the maximum value to the end. The signal descriptors characterizing the amplitude variation are defined from this interpolation process (Figure 4). They correspond to the slopes of the two curves, α_1 and α_2, and to the envelope fluctuations around these two curves, $Err1$ and $Err2$ (in practice, these fluctuations are quantified by the interpolation quadratic error). Even for highly chaotic sounds, these parameters describe a reliable evolution of the sound level and only in a few cases the interpolation was impossible to calculate.

3.2 Evaluation of Descriptors

The 32 signal descriptors defined in the previous section were computed for each sound. The abbreviations chosen for these descriptors are listed in Appendix A. Note that in some cases, the computation leads to aberrant values, which do not reflect the actual sound behaviour. For instance, the calculation of periodicity for sounds, which do not contain cyclic behaviours, would be inaccurate. In such cases, we arbitrarily fixed these meaningless values to 0.
We further aim at determining the most relevant descriptors for each sound category defined previously. This issue can be linked to the one addressed in the field of "data mining" and particularly "feature selection" research. Indeed, feature selection methods aim at optimizing prediction models based on the most explicative parameters. Consequently, they lead to reduce an initial database to its most relevant elements without reducing the performance of the prediction models. This selection is based on criteria which are different across methods and the relevance of features is highly dependent on the definition given to relevance. Hence, a general criterion is: If the exclusion of a feature from the data

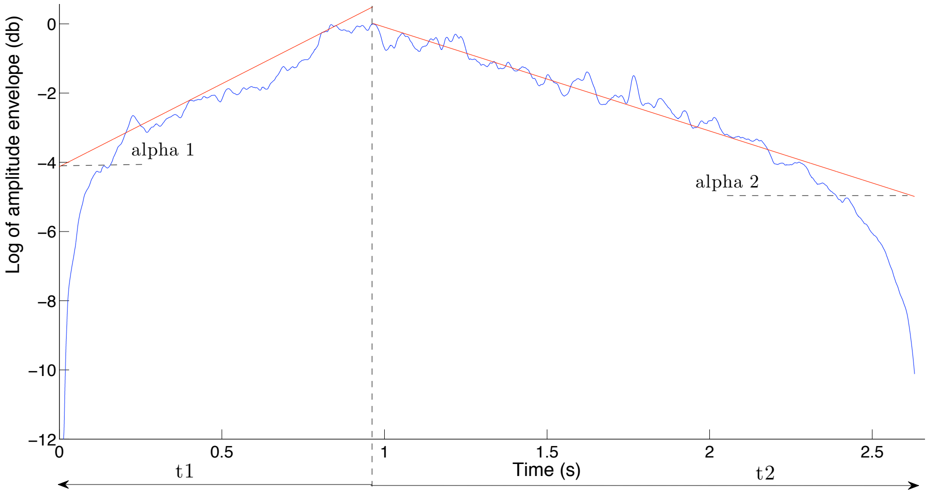

Fig. 4. Example of the Characterization of level variation for a sound from the category "go past" by the 2 slopes α_1 and α_2 of the interpolation process. The duration t_1 (from the beginning of the signal to its max value) and t_2 (from the max to the end) were also considered.

set involves a decrease in the learning model performance, then the excluded feature is defined as relevant. For instance, Blum and Langley [3] give five definitions for "relevant features" and we chose the more "physical" one, which can be summarized as: a relevant descriptor permits differentiation of two different observations (definition 1 in [3]: Relevant to the target).

Choice of the Method: In this study, we do not aim at defining a model which explains sound categories. Indeed, as discussed in section 3 (introduction), model validity depends on the size of the data bank and confidence in categories. Indeed, the set of typical sounds for each category are unequal in terms of number of samples and of portion of subjects that defined them as typical. Furthermore, no perceptual or physical consideration allows us to settle between linear regression and non-linear methods (like regression tree). Such consideration prevents us from building a model. Thus we will use feature selection algorithms that are not linked to a machine learning method (the so called *filter* in opposition with *wrappers*) to determine the most relevant descriptors for each sound category. In particular, the purpose is to build a subset of non-correlated descriptors (to reduce the feature set). In our case, a lot of descriptors are highly correlated, but this can be inline with our definition of relevance. Moreover, from a "physical" point of view, correlations between features can lead to interesting results.
For instance, we can consider correlation between features and classes or consider signal descriptors as statistical distributions, which can be compared with

observations (categories).

Many *filter* methods are available and the results obtained with these various methods can be extremely different. For example, Herrera et al. (2002) [12] compared two different methods[3] on categories of drum sounds and showed that for each category, only one or two features coincided. This difference is due to the fact that one method considers a subset of features and compares results obtained for each subset of a given size, while the other considers each feature independently (and does not consider correlation between features). Otherwise, some methods aim at determining feature combinations (not only linear) that can be more "relevant" than independent features. As discussed in [10], "Two variables that are useless by themselves can be useful together".
For methods necessitating discrete variables, continuous variables are transformed into binary values with an optimized threshold defined from observations. This transformation is useful since different perceptual attributes correspond to different ranges of values of a same continuous physical variable. The borders (threshold values) are generally determined from discrimination tasks. For example, psychoacousticians define both fluctuation strength and roughness from the temporal variations of loudness. Fluctuation strength is defined under 20Hz and roughness between 20Hz and 200Hz. Methods that consider discrete values for a feature match this observation and particularly with Fayyad and Irani's method[4] which takes into account categorization data to perform continuous to discrete transformations. Nevertheless, such methods imply that features are considered independently.
In practice the discretization process is based on binary values and implies to split descriptor values into distinct categories. In our case, the sound categories were defined from a consensus across subjects and consequently the perception is assumed to be non categorical. Considering an intermediate domain between 2 categories would be intuitively accurate (i.e. values under threshold 1, between threshold 1 and 1, above threshold 2).

To summarize, the literature offers a large variety of statistical methods. In our case, we first consider discretization and validity according to the Fisher test[5] and the equivalent for continuous data, the so called "Fisher Filtering" method implemented in R. Rakotomalala's free software TANAGRA[6] (see [10] and [3] for more details on such methods).

Results and Discussion The most relevant signal descriptors highlighted by the 2 methods are presented in Tab. 1 (for discretization method) and in Tab.

[3] Correlation based Feature Selection (CFS) and RefiefF [11]

[4] (1993) cited by [11]

[5] Due to the small number of sounds in each category (from 12 for "rotate" to 8 for "fall"), Fisher Test is more adapted than Chi2 to compare categories and features.

[6] http://eric.univ-lyon2.fr/ ricco/tanagra/fr/tanagra.html

2 (for the fisher filtering method). To test robustness of both methods, we also added to the feature set, a random variable between 0 and 1 attributed to each sound; The results showed that this variable has never been highlighted by the 2 methods and consequently, confirm the confidence in the one presented in this section.

Table 1. Feature discretized according to category and corresponding thresold. [+] means feature value of sound in the category is above the cut-off ([-] under). *Abbreviations are described in appendix A*

Category	”Rotate”	Category: ”Pass by”		Category: ”Fall down”	
Feature	Cut-off	Feature	Cut-off	Feature	Cut-off
cgs_std	224 (Hz) [-]	**Npeak**	3,5 [-]	t_2	0,875[-]
$a0/a1$	0,1 [+]	$\mathbf{Err_2}$	0,26 [-]	**ls**	0,94[-]
$f1$	0,69 (Hz) [-]			α_1	6,1 (db/s) [+]
a1	0,06 [+]			$\mathbf{t_1}$	0,36(s) [-]
f0/f1	0,05 [+]				

Table 2. Signal descriptors and corresponding F-value and P-value obtained by ”Fisher filtering” feature selection method

”Rotate”			”Pass by”			”Fall down”		
feature	F	P-Value	feature	F	P-Value	feature	F	P-Value
a1	19,94	0,000076	$\mathbf{Err_2}$	8,03	0,00748	**ls**	14,60	0,0005
$Npeak$	8,32	0,00658	**Npeak**	7,54	0,00934	Err_1	14,38	0,00055
f0/f1	8,23	0,00684	Err_1	6,31	0,01663	$a1$	9,98	0,0032
			En_mn	4	0,0531	$\mathbf{t_1}$	9,67	0,0036

As expected, results given by ”Fisher Filtering” are coherent with the discretization method. Moreover, they are also coherent with our assumptions. We now discuss results obtained for the motion categories ”Rotate”, ”Pass by” and ”Fall down”. In particular, we focus on the descriptors which have been highlighted by both methods in each category. For the category ”Approach”, no significant descriptors were found by both methods. For the category ”Rise”, only the discretization method highlighted the variation rate of the energy envelope (En_vr) with no statistical criterion for the validity of this descriptor regarding to the category. Note that categories ”Rise” and ”Approach” are the ones with the smallest amount of representative sounds (resp. 4 and 6 sounds selected by more than 50% of the subjects).

Category "Rotate" From a physical point of view, the parameter $a1$ (amplitude of highest peak in the adjusted autocorrelation) corresponds to the most meaningful descriptors since it quantifies the amplitude modulation rate. Note that this parameter is the most relevant (and the only significant) according to Fisher F criterion. Discretization also pointed out this feature with the information that this modulation rate must be over 0,06 (in arbitrary units). The ratio of frequency modulation ($f0/f1$) was also highlighted and indicates that 2 amplitude modulation components are necessary to completely evoke a rotating motion (with no proof according to F value).

Category "Pass by" The relevancy of the Err_2 feature can be explained from physical considerations related to the raise/decay phenomena. Indeed, the decreasing part of the amplitude envelope of the sounds should be log-linear to characterize the "pass by" motion (0.26 [-]in Table1). Err_2 also quantifies smoothness (low fluctuations) of the amplitude envelope, which can explain why it is not significant according to the F criterion (but more correlated with the category). On the contrary, the interpretation of the $Npeak$ feature is more speculative. For the moment, no physical or signal considerations made on the category can be directly related to this parameter.

Category "Fall down" By preliminary listening to the sounds representative of this category, we noticed that they mainly correspond to short impact sounds. This observation is inline with the characteristic of the signal features highlighted by statistical analysis, i.e. short sounds (signal length ls with 0.94 [-]) with an abrupt attack (reflected by parameter t_1 with 0.36 [-]).

These results are of interest from a cognitive point of view since we can deduce that fundamental cues inducing the evocation of falling down motion are not directly contained in sounds. In particular, we can assume that participants associated this motion to sounds by imagining the motion which could have caused the resulting sounds (for instance, striking the floor at the end of its trajectory).

This particular example illustrates the necessity of taking into account some cognitive assumptions additionally to physical considerations.

4 Conclusion and Perspectives

In this paper, we aimed at proposing a global methodology for the design of synthesis tools controlled by high-level parameters, such as mental evocations induced by sounds. Through the particular case of the evocation of motions, this study addressed the general problem of semiotics of sounds for synthesis applications.

The proposed methodology addressed 3 main questions: What are the different categories of motion? What are the common acoustic features of sounds in a category? How to synthesize sounds that evoke specific motions?

First, to determine these different sound categories, we conducted a two-part listening test. The choice of the stimuli was a crucial issue since it constituted the starting point of our methodology. To help subjects to focus on "sensations" evoked by sounds and to avoid bias introduced by the identification of the sound source, we gathered "concrete" samples issued from electro-acoustic music compositions. The first part of the listening test consisted in a free categorization task in which listeners were asked to group sounds as a function of the evoked motions. Groups were quite consistent across subjects and the most representative ones were used in the second part of the listening test, as predefined categories in a constrained categorization task. The predefined categories were represented by prototypical sounds (defined by the previous free categorization test) instead of labels. In this manner, the prototypical sounds imposed an acoustic reference of a given motion category and consequently, listeners assumed to base their strategy mainly on "analytical" properties of sounds. Thus, results allowed determining a set of representative sounds for each category of motion.

Concerning the determination of acoustic features characterizing each sound category, we investigated most of well-known signal descriptors (as defined in mpeg7 standard [17] for example). In particular, according to the wide variety of our sounds (noisiness, complex temporal evolution, different durations...), we compared their evolution by calculating a scalar estimated from time dependent descriptor values (calculation based on successive frames for spectral descriptors). We determined the most relevant descriptors for each category by using the *filter* feature selection method. This method was chosen according to the validity of statistical tests and considerations on the perception of sounds, the results constitute a first step towards the determination of the so-called acoustic invariants, for which different statistical methods have to be examined among hundreds of available methods.

The third main question concerning the building of synthesis tools is still in progress. In particular, the calibration (definition of a valid range of values) of these most relevant descriptors and their control (sound manipulation from the variation of synthesis parameters) are currently being investigated.

Even if many interesting perspectives have been highlighted, we completed the first two steps of our general methodology. Now, we can address with precise questions, each concerned research field such as music, physics, data-mining and cognitive sciences.

Acknowledgments This project has partly been supported by the French National Research Agency (ANR, JC05-41996, "senSons") to Sølvi Ystad.
(http://www.sensons.cnrs-mrs.fr/)

References

1. Aramaki M., Bailleres, H., Brancheriau L., Kronland-Martinet, R., Ystad, S.: Sound Quality Assessment of Wood for Xylophone Bars. Journal of the Acoustical .Society of America, Vol. 121(4), pp 2407-2420, (2007)

2. Bigand, E.: Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition and emotion, 19(8), 1113-1139, Psychology Press, (2005)
3. Blum, A., Langley, P.: Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence, vol. 97(1-2), pp. 245-271, (1997)
4. Defreville, P., Roy, B., Pachet, F.: Automatic Recognition of Urban Sound Sources. Proceedings of the 120th AES Conference, (2006)
5. Eitan, Z., Granot, R. Y.: How music moves: Musical parameters and listeners' images of motion. Music perception, vol. 23(3), 221-247, (2006)
6. Gaver, W. W.: What in the world do we hear? An ecological approach to auditory event perception. Ecological Psychology, 5(1), 1-29, (1993)
7. Gibson, J.J.: The ecological approach to visual perception. Boston: Houghton Mifflin, (1979)
8. Grey, J. M., Gordon, J. W.: Perceptual effects of spectral modifications on musical timbres. The Journal of the Acoustical Society of America, 63(5), pp. 1493-1500, (1978)
9. Guastavino, C.: Categorization of environmental sounds. Canadian Journal of Experimental Psychology, 61(1), pp. 54-63, (2007)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 11571182. (2003)
11. Hall, M.: Correlation-based feature selection of discrete and numeric class machine learning. In Proceedings of the International Conference on Machine Learning, pages 359-366, San Francisco, CA. Morgan Kaufmann Publishers, (2000)
12. Herrera, P., Yeterian, A., Gouyon, F.: Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. Proceedings of Second International Conference on Music and Artificial Intelligence, Edinburgh, Scotland, (2002)
13. Jekosch, U: Assigning Meaning to Sounds - Semiotics in the Context of Product-Sound Design. In Blauert, J. (Ed.), Communication Acoustics, Springer (2005)
14. Kawai, K., Kojima, K., Hirate, K., Yasuoka, M.: Personal evaluation structure of environmental sounds: experiment of subjective evaluation using subjects' own terms" *Journal of sound and vibrations*, (2004)
15. Lufti, A., Wang, W.: Correlational analysis of acoustic cues for the discrimination of auditory motion. Journal of the Acoustical .Society of America, vol. 106(2), August, (1999)
16. McAdams, S.: Recognition of sound sources and events. in McAdams, S. and Bigand, E. (Eds.), Thinking in sound The cognitive psychology of human audition. Oxford University Press, pp. 146-198, (1993)
17. Kim, H. G., Moreau, N., Sikora, T.: MPEG-7 Audio and Beyond: audio content indexing and retrieval. Wiley, (2005)
18. Peeters, G.: A large set audio features for sound description (similarity and classification) in the CUIDADO project. IRCAM, (2004)
19. Portnoff, M. R.: Implementation of the digital phase vocoder using the fast Fourier transform. IEEE Transactions on acoustics, speech and signal processing, vol. 24(3), (1976)
20. Pachet, F., Roy, P.: Exploring billions of audio features Proceedings of CBMI 07, (2007)
21. Schaeffer, P.: Traité des objets musicaux Editions du seuil, (1966)

22. Widmer, G., Dixon, S., Knees, P., Pampalk, E., Pohle, E.: ”From Sound to ”Sense” via Feature Extraction and Machine Learning: Deriving High-level Descriptors for Characterising Music” in P. Polotti and D. Rocchesso (Eds) ”Sound to Sense, Sense to Sound: A State-of-the-Art” (2007)
23. Ystad, S., Kronland-Martinet, R., Schön, D., Besson, M.: Vers une approche acoustique et cognitive de la sémiotique des objets sonores, UST: Théorie et Applications, (2005)
24. Zwicker, E., Fastl, H.: Psycho-acoustics, facts and models. Springer Verlag, (1990)

A Appendix: Abbreviation used for signal descriptors

cgs_std: Standard deviation of spectral centroid
ls: signal length
$f1$: second amplitude modulation frequency
$f0/f1$: ratio of amplitude modulation frequency
$a0$: amplitude modulation ”rate” (first component)
$a1$: amplitude modulation ”rate” (second component)
$a0/a1$: ratio of amplitude modulation ”rates”
$Npeak$: number of peaks in Fourier transform amplitude envelope autocorrelation between 0 and 20 Hz
t_1: time from beginning to max energy point (cf. figure 4)
t_2: time from max energy point to end (cf. figure 4)
Err_1: Quadratic error for linear interpolation of logarithm of amplitude envelope from beginning to max energy point of signal.
Err_2: idem from max energy point to end of signal
α_1: slope of linear interpolation of logarithm of amplitude envelope from beginning to max energy point of signal.(cf. figure 4)
α_2: slope of linear interpolation of logarithm of amplitude envelope from max energy point to end of signal.(cf. figure 4)
En_vr: Variation rate of energy envelope
En_mn: Energy monotonousness