Environmental sound perception: meta-description and modeling based on independent primary studies

Nicolas Misdariis [(1)], Antoine Minard [(1)], Patrick Susini [(1)],
Guillaume Lemaitre [(1)], Stephen McAdams [(2)], Etienne Parizet [(3)]

[(1)] STMS-Ircam-CNRS, 1 place Igor Stravinsky F-75004 Paris, France

[(2)] CIRMMT, Schulich School of Music, McGill University, 555 Sherbrooke St. W. Montreal, QC, Canada H3A 1E3

[(3)] LVA-Insa Lyon, 25bis Avenue Jean Capelle F69621 Villeurbanne Cedex, France

[*misdarii@ircam.fr, smc@music.mcgill.ca, etienne.parizet@insalyon.fr*]

Abstract

The aim of the study is to transpose and extend to a set of environmental sounds the notion of sound descriptors usually used for musical sounds. Four separate primary studies dealing with interior car sounds, air-conditioning units, car horns and closing car doors are considered collectively. The corpus formed by these initial stimuli is submitted to new experimental studies and analyses, both for revealing meta-categories and for defining more precisely the limits of each of the resulting categories. In a second step, the new structure is modeled: common and specific dimensions within each category are derived from the initial results and new investigations of audio features are performed. Furthermore, an automatic classifier based on two audio descriptors and a multinomial logistic regression procedure is implemented and validated with the corpus.

Keywords

environmental sounds, perception, perceptual space, acoustic features, perceptual validation, automatic classification.

Introduction

The purpose of this study is to transpose and extend the timbre description principles of musical sounds to environmental sounds considered, by nature, as non-musical. More precisely, environmental sounds were first defined by Vanderveer [1] as "... any possible audible acoustic event which is caused by motions in the ordinary human environment. (...) Besides 1) having real events as their sources (...), 2) [they] are usually more 'complex' than laboratory sinusoids, (...), 3) [they] are meaningful, in the sense that they specify events in the environment. (...), 4) the sounds to be considered are not part of a communication system, or communication sounds, they are taken in their literal rather than signal or symbolic interpretation."

Within the restricted framework given by the scope of the primary research upon which the present study is based (see Sec. 1), the final aim is also to automate indexing and classification of environmental sounds. This goal is actually essential for sound quality measurements, as well as for further sound-content-based searching and browsing methods that use perceptual models of environmental sounds and often require measurements based on perceptually relevant acoustical similarities. Indeed, in the sound-quality field, most studies use acoustical/psychoacoustic descriptors such as loudness or roughness, in order to explain unpleasantness ratings, whereas several studies have shown that no "universal" descriptors exist for all classes of everyday sounds.

The work detailed in this article starts from four primary industrial studies on sound attributes dealing with sounds produced by car engines (Susini et al. [2, 3, 4], McAdams et al. [5]), air-conditioning units (Susini et al. [6]), car horns (Lemaitre et al. [7, 8]) and closing car doors (Parizet et al. [9]). The aim of these studies was to apply the methodology developed to study the timbre of musical sounds to a specific category of environmental sounds. The standard methodology used in these studies was based on a multidimensional scaling technique (MDS) applied to dissimilarity judgments.

The MDS technique is a fruitful tool for studying perceptual relationships among sounds and for determining the underlying auditory attributes used by participants to rate the perceived similarity among sounds. The term auditory attribute is used to describe the perceived properties or qualities of the sounds. Well-known auditory attributes include loudness, pitch, duration, sharpness, etc. The MDS technique does not require a priori assumptions concerning the number of auditory attributes or their nature, unlike semantic differential methods that use ratings along specific dimensions, such as roughness, for example. The MDS technique represents the perceived similarities in a low-dimensional Euclidean space (referred to as the *perceptual space*), so that the distances among the stimuli reflect the perceived dissimilarities. Each dimension of the space (called a *perceptual dimension*) is assumed to correspond to a perceptual continuum that is common to the whole set of sounds. Thus the main hypothesis with the MDS technique

is that the sounds under study can be compared on auditory attributes that are shared by all sounds in the corpus. In other words, this technique is appropriate for characterizing sounds that are comparable along continuous auditory attributes of a homogenous corpus of sounds, i.e. composed of sounds produced by the same type of source (musical instruments, car sounds, vacuum cleaner noises, etc.). Considering musical sounds, the most common timbre space found by several studies (among which Grey [10], Krumhansl [11], McAdams et al. [12] and Marozeau et al. [13]) consisted of three dimensions correlated with acoustic features in order to associate a measurable sound parameter with each perceptual dimension of timbre. The assumption of this approach rests on the model suggested by McAdams [14], who postulates that the recognition of the sound sources arises from a process of analysis, computation and extraction of a certain number of auditory features related to the acoustic parameters of the signals. Then, in many of these musical timbre studies, the three dimensions were found to be significantly correlated with a spectral feature that most often represented auditory brightness (energy distribution along the frequency scale), a temporal feature that characterized attack and a spectro-temporal feature corresponding to spectral variations over time. The MDS technique has been shown to be an efficient tool for revealing and describing the unknown auditory attributes underlying the timbre of musical sounds.

In the present context, environmental sound studies, experimental data, analyses and acoustic parameters have been reviewed and compared from the four initial studies. An investigation of these combined data was conducted, and an attempt to model the resulting structures on the basis of the primary results was made using generalized toolboxes (essentially, "Ircamdescriptor" from Peeters [15] and "Auditory Toolbox" from Slaney [16]) in order to unify – and in some cases to improve – the description of the initial data. Here we will first introduce and describe all the studies taken into account in this review, their stimulus sets, the experiments performed, the resulting perceptual spaces and the correlated acoustic features. Then, in order to contribute to environmental sound perception, we will first present this overall stimulus set organization in terms of the main environmental sound classes, propose both inter-class and intra-class structure descriptions, and finally initiate an automatic classification modeling approach within the restricted scope of the present study but on the basis of perceptually relevant data and results gathered during its experimental parts.

1. Primary studies

We present in this section the frameworks, motivations and results of the four experimental studies that represent the starting point of our meta-analysis. These studies focus on the sounds from:

A.- Car interiors (Susini et al. [2, 3, 4], McAdams et al. [5])

B.- Interior air-conditioning units (Susini et al. [6])

C.- Car horns (Lemaitre et al. [7, 8])

D.- Closing car doors (Parizet et al. [9])

These four studies all addressed the issue of sound quality and shared a common approach: they studied the timbre of the different types of sounds. More precisely, they use a common methodology and share similar analysis techniques. This procedure relies on the psychoacoustic definition of timbre: "Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar." (American national standard acoustical terminology" (1994). American National Standards Institute, ANSI S1.1-1994 (R1999); see also Krumhansl [11]). Timbre is thought to be multidimensional, encompassing several perceptual attributes that are collectively referred to by this term. In order to uncover the attributes of timbre, the methodology used in the studies was based on the procedure developed to study the timbre of musical sounds (McAdams et al. [12]). It has three main steps, the first one being sometimes preceded by an preliminary step (labeled "0" below) used to reduce the number of sounds to be tested in the first step:

0. Because the following step of the methodology needs a small number of sounds to be experimentally feasible, a preliminary step is sometimes used in order to reduce the original corpus to an acceptable number of stimuli (usually not more than 20 samples). Free-sorting tasks and cluster analyses (see Sec. 2.1 for further details) are used to attain this goal. A free-sorting task consists in asking participants to sort the sounds of the set into as many categories as they wish. Thus, they identify the main categories of sounds that are studied and allow for the selection of representative subsets of sounds by homogeneously sampling across the categories.
1. A dissimilarity rating experiment collects the perceived dissimilarities among the sounds, which are then used as proximity data. It consists in asking the participants to rate directly the dissimilarity between both sounds of each possible pair within the set of sounds. The evaluation is made on a continuous scale labelled "Very Similar" at the left end and "Very Dissimilar" at the right end. It has the great advantage that it does not impose any predefined rating criteria on the listener.
2. The proximity data are modeled with a multidimensional scaling (MDS) analysis that fits distances in a geometrical space to the dissimilarity data. The dimensions of this space represent the perceptual dimensions underlying the proximities. Different levels of complexity exist in the MDS approach depending on the model and associated algorithm (see Appendix A); in the present case, two particular MDS techniques were used in the studies: the INDSCAL (Individual Difference Scaling) and CLASCAL (Latent Class Approach) models.
3. The final step of a timbre study is to give a physical interpretation of the perceptual dimensions revealed by the MDS analysis. This is usually done by submitting the perceptual dimensions to linear regression analyses with relevant acoustic features. Some of them are psychoacoustic

descriptors, i.e. acoustic features that have been found to correspond to auditory sensations. Models that compute psychoacoustic descriptors are usually based on a model of the peripheral auditory system.

1.1 Studies A (A1, A2): car interior [2, 3, 4, 5]

Context

The main goal of this study was to analyze the timbre of the sounds of car interiors in a given driving condition from the driver/passenger point of view.

Stimuli

The sounds were recorded in 16 different vehicles at two different engine modes. The engine modes defined two sub-studies: **study A1** involved sounds produced when the engine was running in 3rd gear at 4000 RPM (Round Per Minute) and **study A2** involved sounds produced when the engine was running in 5th gear at 3500 RPM. A preliminary experiment showed that loudness was the main auditory cue used by the participants to rate the dissimilarity. Thus, in order to let other auditory attributes emerge, loudness was equalized. Both stimulus sets were composed of 16 stereophonic sounds that were 4.1 seconds in duration. Their levels – after loudness equalization – varied between 69 and 80 dB SPL (Sound Pressure Level).

Participants

For each engine mode stimulus set, a dissimilarity rating experiment was conducted with 30 participants.

Analysis and Results for Study A1

A CLASCAL analysis (see Appendix A) of the data yielded a 1-latent class, 3-dimensional space with specificities. Figures B1.1 to B1.3 in Appendix B represent the projections of the space, and Table B1 reports the correlation coefficients of the acoustic features best fitting the perceptual dimensions. The first dimension is correlated [$r(14)=-0.81$, $p<0.01$] with a feature corresponding to the relative balance of the harmonic (motor) and noise (air turbulence) components. The second dimension is correlated [$r(14)=-0.70$, $p<0.01$] with a variation of the spectral centroid with the frequency dimension represented in ERB-rate (see Appendix B for more details). The third dimension is significantly correlated [$r(14)=-0.83$, $p<0.01$] with an acoustic feature quantifying the spectral decrease of the harmonic part of the sound.

Analysis and Results for Study A2

A CLASCAL analysis (see Appendix A) yielded a 1-latent class, 2-dimensional space with specificities. Figure B2.1 in Appendix B represents the perceptual space and Table B2 reports the correlation coefficients of the acoustic features best fitting the perceptual dimensions; the features that are the best correlated with the two dimensions are also reported in Table B2. The first dimension is correlated [$r(12)=0.93$, $p<0.01$] with an acoustic feature conveying the relative balance between two groups of

partials of the the harmonic part of the signal (see Appendix B for more details). The second dimension is correlated [$r(12)=0.86$, $p<0.01$] with the spectral centroid computed on the C-weighted version of the signal (see Appendix B for more details)

1.2 Study B: interior air-conditioning units [6]

Context

This study focused on the sound quality of interior air-conditioning units.

Stimuli

The initial set consisted of 43 sounds produced by units of different brands. A free-sorting experiment was first conducted to select an homogeneous subset of sounds representative of the existing range for this type of sounds. The results of this experiment also showed that three categories were made mainly by grouping together sounds with similar loudness levels. As in study A, in order to prevent loudness from dominating the ratings (possibly masking more subtle effects), the sounds were selected in the category corresponding to a medium loudness level (average level: 46.5 dB SPL, 2.2 dB standard deviation). An informal experiment was then performed with only 5 participants to get an initial estimate of the perceptual space structure. The outcome of the MDS analysis was that the space was not homogeneously sampled. Therefore synthesized sounds were added and redundant sounds were removed in order to produce a more homogeneously distributed stimulus set. The synthesized sounds were created on the basis of features of the sounds in the stimulus set, using a geometric interpolation within the space. The resulting stimulus set consisted of 19 sounds: 15 recordings of air-conditioning units and 4 synthesized sounds. They were all 5.9 seconds in duration with levels varying between 44 and 52 dB SPL.

Participants

The dissimilarity rating experiment was conducted with 50 participants.

Analysis and Results

A CLASCAL analysis (see Appendix A) of the dissimilarity ratings yielded a 5-latent class, 3-dimensional space with specificities. Figures B3.1 to B3.3 in Appendix B represent the projections of the 3-dimensional space, and Table B3 presents the correlation coefficients of the features that are the best correlated with the perceptual dimensions. The first dimension is correlated [$r(17)=-0.97$, $p<0.01$] with a feature corresponding to the relative balance of the harmonic (motor) and noise (air turbulence) components. The second dimension is correlated with a frequency-weighted variation of the spectral centroid of the noise component [$r(17)=0.73$, $p<0.01$]. The third dimension is correlated with loudness [$r(17)=0.84$, $p<0.01$]. Indeed, even though the selected sounds are in the same range of loudness, they were not equalized in loudness.

1.3 Study C: car horns [7,8]

Context

This study concerned the timbre of car horns in order to define specifications for the design of new sounds. The initial stimulus set consisted of 43 recordings of current car horn sounds. These sounds can be either monophonic (one note) or polyphonic (two or three notes to make a chord) and are produced by two different mechanisms: a metal plate or a folded horn that acts as a resonator and is attached to the membrane of an electroacoustic driver. Both produce very specific timbres. A preliminary sorting experiment highlighted 9 main categories of sounds connected with these different mechanisms and properties.

Stimuli

A sample of 22 sounds was chosen among the 9 categories. Among these 22 sounds, 13 were monophonic and 9 were polyphonic, 10 were produced by "plate" resonators and 12 by "horn" resonators. All sounds lasted between 0.6 and 2.2 seconds. Their levels varied between 63 and 77 dB SPL.

Participants

A dissimilarity rating experiment using this set of sounds was conducted with 41 participants.

Analysis and Results

The dissimilarity ratings were submitted to a CLASCAL analysis (see Appendix A), resulting in a 6-latent class, 3-dimensionial space with specificities. Figures B4.1 to B4.3 in Appendix B represent the projections of the 3-dimensional space, and Table B4 reports the best-correlated features (see [7], for more details on the acoustic features). The first dimension is correlated [$r(20)=-0.9$, $p<0.01$] with roughness. The second dimension is correlated [$r(20)=0.9$, $p<0.01$] with a variation of the spectral centroid integrating a perceptual approach to compute this parameter (ERB scale, see Marozeau et al. [13]). The third dimension is correlated [$r(20)=-0.8$, $p<0.01$] with an acoustic feature related to the fine structure of the spectral envelope.

1.4 Study D: car door closing [9]

Context

The main goal was to study the timbre of car door closing sounds in the context of evaluating their sound quality.

Stimuli

An initial set of 27 stereophonic recordings (16 sounds from different cars and 11 sounds from two cars with modified seals) was submitted to a sorting experiment with 31 participants in order to select a representative subset of 12 sounds. Among these 12 sounds, 4 were recorded from cars with modified

seals. The durations of the sounds varied between 0.3 and 0.5 seconds, and their levels varied between 66 and 84 dB SPL.

Participants

A dissimilarity rating experiment was conducted with 40 participants.

Analysis and Results

The dissimilarity data were submitted to an INDSCAL analysis (see Appendix A). A 3-dimensional space was found. Figures B5.1 to B5.3 in Appendix B represent its projections, and Table B5 reports the correlation coefficients of the features best correlated with the perceptual dimensions. The first dimension is correlated with a feature corresponding to sharpness, as defined by Aures [36] [$r(10)=-0.90$, $p<0.01$], as well as to the spectral centroid [$r(10)=-0.93$, $p<0.01$]. The second dimension is correlated [$r(10)=0.87$, $p<0.01$] with an indicator related to the temporal evolution of instantaneous loudness, according to Zwicker's model [20]. No descriptor was found that correlated significantly with the third dimension.

1.5 Comparisons and discussion

The studies reported in the previous subsections identify the perceptual space of sounds contained in five separate stimulus sets (labelled A1, A2, B, C and D), associated with different kinds of environmental situations, mainly related to car and appliance industries. The results of these studies are summarized in Table 1 below.

	A- Car interior	**B- Air-conditionning units**	**C- Car horns**	**D- Car door closing**
Corpus	A1: 16 snds 3rd gear, 4000 rpm A2: 14 snds, 5th gear, 3500 rpm	19 sounds (4 synthesized)	22 sounds	12 sounds
Analysis	CLASCAL	CLASCAL	CLASCAL	INDSCAL
Results	A1: 3 dim., specif., 1 lat. class. A2: 2 dim., specif., 1 lat. class.	3 dim., specif., 5 lat. class.	3 dim., specif., 6 lat. class.	3 dim.
Descriptors	A1, dim.1: RAPmv-A A1, dim.2: CGg-ERB A1, dim.3: Dec A2, dim.1: rad_2N/0.5N A2, dim.2: CGg-C	dim.1: NHR-A dim.2: Sc_n-B dim.3: N (Loudness)	dim.1: Roughness dim.2: Spectral centroid dim.3: Spectral deviation	dim.1: Spectral centroid dim.2:Cleanness indicator dim.3: …

Table 1: Table recapitulating methodological context and main results for studies A1, A2, B, C, and D (see corresponding sections above and Appendix B for further details)

A comparison of the acoustic features correlated with the dimensions of these four perceptual spaces yields some interesting facts:

- In every study, at least one dimension in the perceptual space resulting from the MDS analysis is found to be related to a spectral centroid feature, usually describing the "brightness" of a sound. This "brightness" attribute seems therefore incontrovertible when trying to compare two sounds of any of these kinds of sources. However, this attribute seems to take different forms acoss the studies:
    - It can be computed with a frequency weighting representing the variation in sensitivity of the human ear over the audible frequency range at different presentation levels (A-, B- and C-weightings).
    - It can introduce a much more sophisticated model of the hearing process (ERB filters).
    - It is sometimes only computed on a particular part of the signal (noise part).

  The subsidiary questions are: will all these "brightness" predictors be as efficient for all studies? If not, is there a particular calculation that fits all of the spaces equally well?
- In 3 studies, relevant acoustic features appear to include separate calculations for the harmonic and noise parts of the signals. The signal processing needed for this separation is quite complex and often includes the setting of several initial algorithm parameters. Again, the question raised is: will a common set of these parameters result in the same efficiency of separation for the correlation scores in the 3 studies?
- For the 2 other studies, the correlation results exhibit specific relevant acoustic features. This fact confirms that a universal low-dimensional perceptual space describing all sounds does not exist. It would also tend to agree with McAdams et al [5] and McAdams [21] who observed that when sounds are produced by too different kinds of sources, the dissimilarity judgments may be based on cognitive factors rather than on perceptual signal-related ones, which results in a strongly categorical description.

2. Meta-processing: complementary experimental investigations

The MDS technique is appropriate to characterize a set of sounds caused by very similar sound sources, but not for different and obviously identified sources. For instance, McAdams et al [5] applied an MDS analysis to an extremely heterogeneous set of environmental sounds (trains, cars and planes). The analysis yielded a strongly categorical perceptual structure: listeners identified the sound sources rather than comparing them along continuous perceptual dimensions. In that case, participants based their perception on a predominant cognitive factor: recognition, classification, and identification of the sound source (see McAdams [21]). In other words, when the sounds under consideration are similar, which means that they are provided by the same type of sources, listeners are able to compare them on continous perceptual dimensions, otherwise they are categorized by association with the type of source. As a consequence, the

perceptual organization of the five groups of stimuli may be based on a 2-level structure displaying both categorical and continuous levels (see Figure 1, for illustration):

- a categorical (discrete) level corresponding to the main sound event categories, each of them being related to a distinct physical cause or source;
- a continuous level that will associate each of these categories with a perceptual space with salient dimensions that can be either specific to the given category or shared with the others.

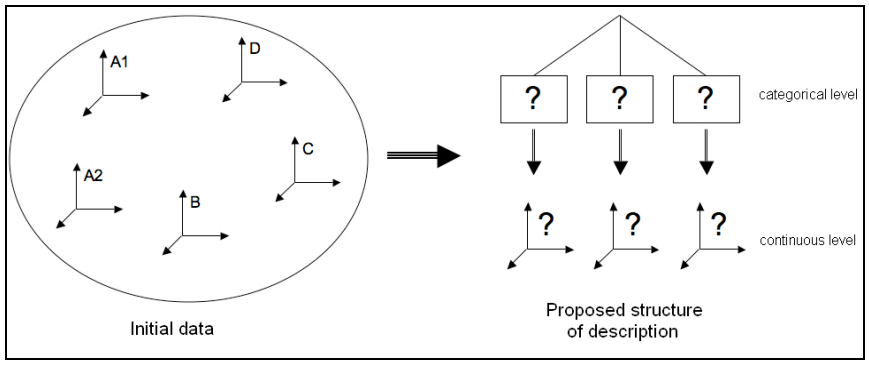

Figure 1: Schematic representation of the proposed 2-level organisation structure.

In order to evaluate the consistency of this structure and to validate it within the scope of the present research, an additional experimental investigation was conducted consisting of two successive experiments:

1. a free-sorting task to identify the main sound event categories composing the overall sound corpus, combined from the initial corpora presented in Sec. 1;
2. a forced-choice sorting task on more heterogeneous sounds (extracted from commercial sound librairies) in order to extend and determine more precisely the boundaries of these categories.

The results of these experiments will then be used in the last part of the study (see Sec. 3) in order to define new ways of modeling the structure on both discrete and continuous levels, as defined in our hypothesis, that describes the main sound event categories and perceptual dimensions attached to each of those categories, respectively.

2.1 Experiment: Free sorting task on the initial corpora

In order to identify the main perceptual categories among the sounds under consideration in this study, a free-sorting experiment with this complete stimulus set was conducted.

Method

***Participants*.** Twenty participants (8 women and 12 men) volunteered as listeners for this

experiment and were paid for their participation. All reported having normal hearing.

Stimuli. The resulting unified stimulus set is a collection of 83 sounds distributed as follows: 16, 14, 19, 22 and 12 sounds from studies A1, A2, B, C and D, respectively. See the *Stimuli* subsections of Sec. 1 for more details on the sounds. In order to prevent the listeners from sorting the sounds according to their loudnesses, a preliminary loudness-equalization experiment was conducted with 7 participants working at IRCAM, resulting in an 83-sound loudness-equalized corpus.

Apparatus. Testing took place in a double-walled IAC sound-isolation booth. The sounds were played over Sennheiser HD 520 II headphones through a RME Fireface 400 audio card plugged into a Macintosh Mac Pro (Mac OS X v10.4 Tiger) workstation. The test was run using a Graphical User Interface (GUI) specifically developped in Matlab (v7.0.4) including stimulus control, data recording and sound play-back[1].

Procedure. At the beginning of the procedure the participants were given written instructions briefly presenting the context of the study and detailing the task to be performed. The task was to classify the 83 sounds of the corpus in as many categories as they wished according to their own criteria and, in a second step, to select the most representative sound – the prototype – for each of the classes (see Sec. 2.2 for the definition of a prototype by Rosch [24]). In the GUI, the sounds were represented as dots that could be either played (double-click) or moved (drag and drop) in the dedicated area of the screen in order to graphically compose the categories (see Figure C1 in Appendix C for an illustration of the interface).

Results

Analysis. The experimental data consist in individual incidence matrices (coding the set partitions of each subject) that are summed to form a co-occurrence matrix. The co-occurrence matrix represents how many subjects have placed each pair of sounds in the same category. This can also be interpreted as a proximity matrix (Kruskal [17]). In the present case, we derived a hierarchical tree representation from these data using an unweighted arithmetic average clustering (UPGMA) analysis algorithm (see Legendre et al. [22] for computational details). In such a representation, the distance between two sounds is represented by the height of the node which links them. Among the 91,881 triplets that can be formed out of 83 sounds, 94% follow the ultrametric inequality, which shows the adequacy of the tree representation for these data (see Legendre et al. [22]). The tree representation is shown in Figure 2. It can be clearly seen that 3 main categories constitute the unified corpus. Looking in detail at the items inside each of them, we can observe that these 3 categories correspond respectively to studies A and B (right part of Fig. 2), study C (left part of Fig. 2) and study D (middle part of Fig. 2). Moreover, listening to these items led us to propose a semantic labeling for each of these 3 categories: "motor", "instrument-like", and

[1] This GUI was developped by Vincent Rioux.

"impact", respectively.

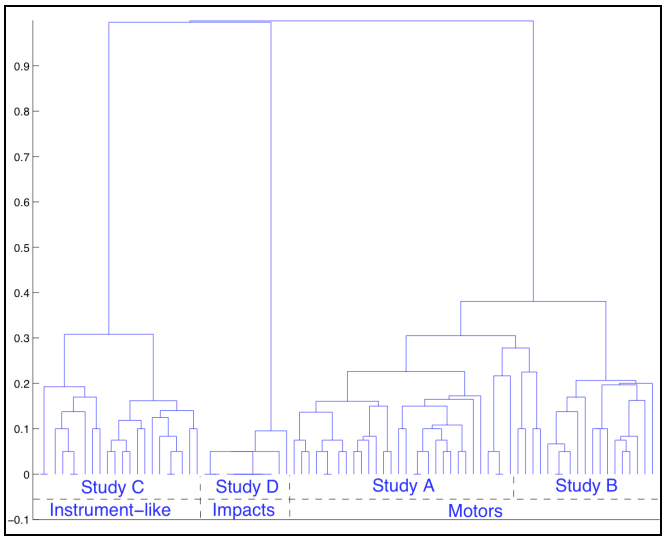

Figure 2: Experiment 1, dendrogram resulting from the cluster analysis - representation of the 3 main categories: *motor* (right part), *instrument-like* (left part) and *impact* (middle part).

We subsequently extracted the prototypic sound for each of the 3 categories by a specifically developed algorithm. Each listener selected prototypes with regard to her/his own categories, which are not necessarily the 3 categories extracted from the cluster analysis. Consequently, we had to consider the prototype selection for each pair of sounds, as follows:

- In an individual 83 X 83 matrix, for each pair of sounds, if the sound *j* was selected by a listener as prototype of a category that contains a sound *i*, then the cell *(i,j)* is incremented (but not the cell *(j,i)*).
- After summing the matrix over the panel of listeners, a sub-matrix is extracted, for each of the 3 final categories with the rows and lines indexed by the sounds constituting the category.
- Each obtained sub-matrix is averaged over its rows, and the highest score gives the index of the prototype.

With this method, the selection by a listener of a sound as prototype for sounds that do not belong to the same final category does not influence the final selection of prototypes. The 3 selected sounds will be used in Experiment 2 (Sec. 2.2) as a definition of the 3 categories.

Discussion. As a result, the five initial corpora can be reorganized into 3 main categories on the basis of the perceptual results (Experiment 1). Obviously, these categories are strongly defined by the initial studies from which they were drawn. In other words, there is no overlap between the initial corpora and the final structure: studies A1, A2, B belong to a first category, study C to a second category and

study D to a third one. However, this fact was intuited before the experiment just by listening to the sounds. According to the sound production type, we semantically defined these categories as:

• "**motor"** (first category): sounds from both car interior (studies A1 and A2) and air-conditioning unit (study B) corpora. These sounds have two discriminable components: a harmonic part with a quite low fundamental frequency produced by a "motor" and a noisy part produced by air turbulence.

• "**instrument-like"** (second category): sounds that correspond to the car horn corpus (study C), which are defined by one or several higher tones, closer to those produced by musical instruments than those generated by motors.

• "**impact**" (third category): sounds of the car door closing corpus (study D). Actually, one can easily discriminate these sounds from the others because of their temporal structure. This idea is consistent with the discrimination of percussive and sustained sounds among musical timbres. Indeed, impact sounds of the environment are quite close to musical percussive sounds in terms of sound production.

This categorization is consistent with the product sound classification proposed by Özcan et al. [23] defined by 6 sound categories: *air, alarm, cyclic, impact, liquid* and *mechanical*. Even though these product sounds were from a domestic context, Özcan et al. found an ***impact*** category; they also found an *alarm* category that can correspond to the present ***instrument-like*** category with regard to basic similarities in pitch, harmonic structure or stationary aspects of the sounds; and finally, the present ***motor*** category can be linked to both their *air* and *mechanical* categories.

2.2 Experiment 2: Forced-choice sorting task on an extended corpus

On the basis of the previous results (3 main categories of environmental sounds within the scope of the unified corpus, with an associated prototype for each), a second experiment was conducted in order to generate a more heterogeneous corpus that would better represent the range of variation of each category. This was done by means of a forced-choice procedure, the main choices being the categories found in Experiment 1. These 3 categories were each identified by their respective prototypic sound extracted from Experiment 1, instead of being verbally defined as it is usually done in this kind of procedure. The notion of prototype is based on psychological principles related to the way one organizes knowledge of the surrounding world. For Rosch [24], a prototype is the element of a group that is the most similar to all items inside the group and, at the same time, that is on average the most different from all items of all the other groups. The notion of prototype used in the present study is directly derived from Rosch's concept.

Furthermore, the outcome of this second experiment will also provide perceptually validated data for the modeling part of the present study in order to implement an automatic classifier (see Sec. 3.2).

Method

***Participants*.** Twenty-one participants (8 women and 13 men) volunteered as listeners for this experiment and were paid for their participation. All reported having normal hearing.

***Stimuli*.** A new extended corpus was created on the basis of the main categories found in the previous experiment. Several sounds were added to each category in order to make the stimulus set more complete and heterogeneous. We therefore chose various new sounds with quite extreme cases for each category from commercial sound libraries (Hollywood Edge Premiere Edition I, II and III, Sound Ideas General Series 6000 and Blue Box Audio Wav). Here are some examples of sounds added in each category:

- for *motor* sounds: truck, aircraft, motorbike, helicopter, crane, vacuum cleaner, fridge, blender, electric shaver, lawn mower;
- for *instrument-like* sounds: phone ringing, dishes squeak, door creak, alarm, bell;
- for *impact* sounds: glass shock, various doors closing (fridge door, house door, etc.), computer keyboard, water drop, tennis ball.

Again, the sounds needed to be equalized in loudness so that the judgements would not be based on this auditory attribute. However, considering the high number of sounds, a preliminary experiment of loudness equalization would have been quite long. As a consequence, the sounds' loudnesses were equalized with regard to the value given by the loudness model of Zwicker et al. [28].

The final corpus was composed of 150 loudness-equalized sounds with an equal distribution of 50 sounds in each category.

***Apparatus*.** The same technical equipment as in Experiment 1 was used. However, the study was run using a GUI specifically developped in the PsiExp v3.4 experimentation environment including stimulus control and data recording (Smith [25]). The sounds were played with Cycling ’74’s Max/MSP software (v4.6).

***Procedure*.** At the beginning of the experiment, the participants were given written instructions briefly presenting the context of the study and detailing the task to be performed. They were asked to classify the 150 sounds of the new corpus into 3 unnamed categories associated with their respective prototypical sounds by clicking on the corresponding button. A fourth button labeled "other" allowed participants to not choose any of the 3 main categories (see Figure C2 in Appendix C for an illustration of the interface). The specifity of the present paradigm was to make the categories explicit with the prototype sounds found in Experiment 1 – with the obvious exception of the class "other" – instead of naming them directly. This implementation was chosen in order to avoid any ambiguity in the understanding of the arbitrary semantic attributes that did not result from verbalization analyses.

Results

Analysis. Table 2 presents the sound distribution, i.e. mean and standard deviation of the number of sounds placed by the participants in each category. Note that these data are strongly influenced by the choices of sounds added by the experimenter, and that these numbers mainly show the adequacy or inadequacy of these choices. However, the following points may be emphasized:

• The high standard deviation of the number of rejected sounds ("other") might be related to differences in strategy among the participants who did not use the same selectivity threshold, or the same granularity, to group the sounds.

• The high mean number of sounds combined with a relatively low standard deviation for the motors shows a consensus among the participants that proves the adequacy of the selection of sounds for this category.

• On the contrary, the relatively high standard deviations for the two other categories show some variability in the listeners' judgements, which is probably due to the quite large variety of chosen sounds for these categories. For the "instrument-like" category, the variability seems to be related to the difficulty of theoretically defining this type of sound, whereas for the impacts, it could be explained by the too-general character of this category.

Nevertheless, after computing a percentage of belonging to the categories for each sound of the 150-item corpus, we observed that these disparities in classification were concentrated on certain sounds only, which were then rejected (26 sounds under a threshold of about 65% of belonging to a category). We thus obtained a selection of the extended corpus leading to a 124-sound stimulus set: 50 "motor", 27 "instrument-like", 47 "impact". Note that this final distribution corresponds roughly to that of Tab. 2.

	Prototype #1 (*motor*)	Prototype #2 (*instrum-like*)	Prototype #3 (*impact*)	"other"
mean	48.7	32.6	45.9	22.7
std	5.5	8.3	13.2	15.5

Table 2: Experiment 2 – distribution of the sounds in the experimental categories.

Discussion. The partitioning of the data across the 3 categories shows a good consensus on a certain number of sounds for each class. With this result, we are then able to make a selection of sounds that are clearly associated with one of the 3 categories revealed in Experiment 1. This leads to the constitution of a perceptually validated sound corpus with regard to the *motor*, *intrument-like* and *impact* categories, which is now large and representative enough to consider the conception and validation of a predictive tool for automatic classification of environmental sounds in these three categories.

2.3 Discussion

Within the restricted scope of environmental sounds studied here (industrial sounds from cars and machines), we are now faced with the following structure:

- a *motor* category including 49 sounds from 3 different corpora (A1, A2, B), each of them being described by a perceptual space and augmented with 50 perceptually validated new sounds, for a total of 99 items,
- an *instrument-like* category including 22 sounds from corpus C described by a perceptual space and augmented with 27 perceptually validated new sounds, for a total of 49,
- an *impact* category including 12 sounds from corpus D described by a perceptual space and augmented with 47 perceptually validated new sounds, for a total of 59.

In the next step, this corpus will serve as input for the implementation of the automatic classifier detailed in Sec. 3.2.

3. Meta-processing: modeling the description structure

This section was designed to confirm the second part of the starting hypothesis, which stipulates that both inter-category and intra-category properties exist, i.e. dimensions shared by the all categories and specific dimensions related to their mutual discriminating differences. Furthermore, the knowledge of these discriminating features could facilitate the implementation of a predictive tool capable of automatically recognizing whether a new item belongs to one of the 3 meta-categories. Note that every acoustic feature mentioned in this section is extracted either from the Ircamdescriptor toolbox (CUIDADO project, Peeters [15]) or from the Auditory Toolbox (Slaney [16]).

3.1 Continuous level: unifying the perceptual space dimensions

In this first part, we investigate the shared and specific properties across the categories by considering the data coming from the corpus described by perceptual space dimensions (corpuses A to D). The main idea is to unify these data by recomputing the acoustic features explaining the different perceptual dimensions in a more systematic manner, in order to point out regularities and singularities among the given spaces. The implementation of these acoustic features is detailed in Appendix D.

Note that some of the stimulus sets contain only monophonic sounds, whereas others contain only stereophonic sounds, and, although the acoustic features are calculated on both channels in the latter case, the salience of an indicator in one channel compared to the other depends on the recording context. For example, if a car interior sound has been recorded from the driver's seat, the most relevant channel for a given sound feature will probably not be the same as if it had been recorded from the passenger's seat.

Accordingly, the features in the correlation tables can be either from the left or the right channel, or from the mean of both channels.

3.1.1 MDS analyses compatibility

The two models giving rise to the perceptual spaces that will be unified in this section are INDSCAL and CLASCAL (see Appendix A). As both models remove the rotational invariance of the obtained spaces, one could assume that both models would result in similar main perceptual dimensions (even if possible slight differences on items' position or axes's orientation may be due to the precision of the model). However the presence of specificities in the latter can modify the psychological meaning of the dimensions. Indeed, the fact that a part of the Euclidean distances is explained by those specificities leads to a modification of the proportion explained by the dimensions. Thus the dimensions obtained by both models will not necessarily be the same.

All the same, the only sound corpus for which the INDSCAL method was used (study D) corresponded to a different sound category than those of the other corpora (see Sec. 2.1). As a consequence, the fact that the dimensions were obtained differently from the other studies is not a problem. Indeed, this perceptual space will be studied separately from the others.

3.1.2 *Motor* category

One of the main characteristic of this kind of sound is that it contains two different simultaneous parts. The first one corresponds to a harmonic pattern that can be easily modeled by a sum of sinusoids, and the second one corresponds to the noise resulting from the air turbulence. Perceptually, these two parts are highly discriminable. Consequently, unlike the other two categories, both parts need to be taken into account independently when estimating the acoustic features This is the reason why harmonic separation methods were tested and used in order to describe both parts, as well as their mutual interaction.

This meta-category regroups stimulus sets A1, A2 and B, presented in Secs. 1.1 and 1.2. Those stimulus sets' MDS analyses resulted in 3-dimensional perceptual spaces, except for that of study A2, which gave a 2-dimensional space. Because of the relative proximity of the sounds coming from these 3 stimulus sets, two shared dimensions were found. The first one is related to the harmonic/noise ratio, while the second is related to the spectral centroids of both parts with some interactions. Finally, the stimulus sets of studies A1 and B differ in their third dimension (the stimulus set of study A2 only gave a 2-dimensional space), most likely because of a practical particularity of the experimental protocol: the sounds of set A1 were first loudness-equalized, unlike those of set B. The correlation scores between dimensions of the motor sound stimulus sets and the best-fitting acoustic features (see Appendix D) are presented in Tab. 3.

	Dimension 1	Dimension 2	Dimension 3
-- study A1 --			
HNR	**-0.93****	0.12	-0.22
Complex brightness	0.09	**0.86****	-0.17
PSS	0.07	-0.15	**0.83****
-- study A2 --			
HNR	**0.83****	-0.17	
Complex brightness	-0.34	**0.90****	
-- study B ---			
HNR	**0.91****	-0.07	0.47
Complex brightness	-0.52	**0.81****	0.00
Loudness	0.42	-0.07	**0.84****

Table 3: Correlations between acoustic features and dimensions of the *motor* category / studies A1, A2, B (*df* = 14, 12, 17, respectively, ** $p < 0.01$).

• Dimension 1: Harmonic emergence (**HNR**)

For all three stimulus sets, several acoustic features correlate highly with this shared dimension, but they were of quite different types, and not all of them were significant. Furthermore, only one feature correlated well with this first dimension for the three stimulus sets: the Harmonic-to-Noise Ratio (HNR). Perceptual differences in the sounds along this dimension are related to the amount of harmonic (or pseudo-harmonic) energy in the signal. The HNR linear regressions with the first dimension of every motor stimulus set are shown in Figs. 3.1 to 3.3. The other features that correlated highly with this dimension were usually spectral envelope features. Actually, those high correlation scores are consequences of the HNR correlation. Indeed, the spectral envelopes of both parts of the sounds have quite different behaviors, and when the proportion of both parts is modified, the overall spectral aspect of the sound is also modified.

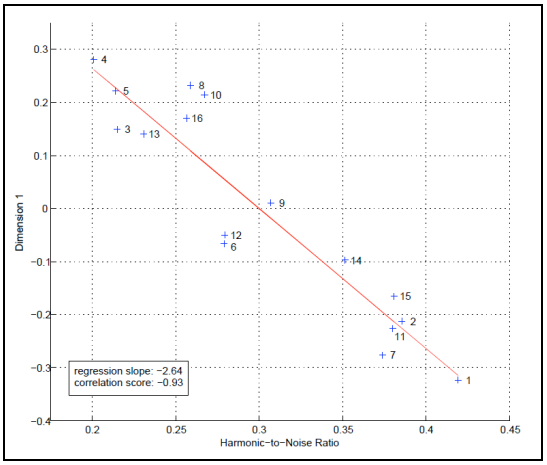

Fig. 3.1: Linear regression between dim. 1 and HNR, *motor* class / study A1.

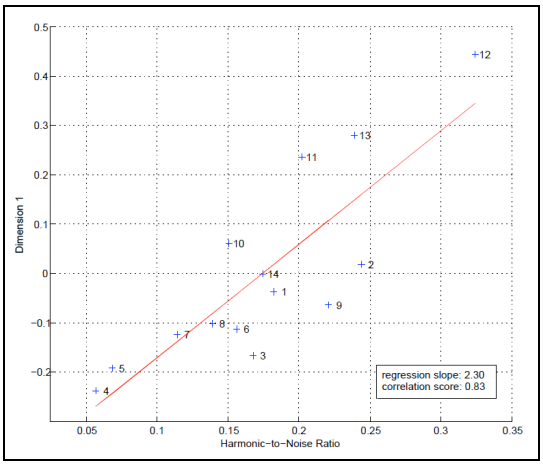

Fig. 3.2: Linear regression between dim. 1 and HNR, *motor* class / study A2.

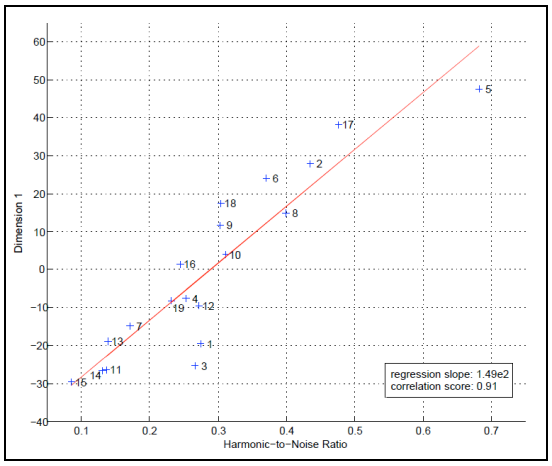

Fig. 3.3: Linear regression between dim. 1 and HNR, *motor* class / study B.

• Dimension 2: ***Complex brightness***

For the three stimulus sets, when listening to the sounds along this scale, brightness features, such as spectral centroid or sharpness, seem to explain the dimension. However, for the two stimulus sets in which the harmonic part is most prevalent, i.e. sets A1 and B, the perception of brightness seems to depend on the harmonic proportion. Indeed, the brightness perception of a predominantly noisy sound is not the same as that of a predominantly harmonic sound, all the more because both parts have quite different spectral behaviors: the energy of the harmonic part is quite concentrated in the low frequencies for this type of sounds. It is thus essential to take into account both the harmonic and noise parts in the brightness estimation. That is the reason why multidimensional linear regression theory (see Legendre et al. [22]) is applied in order to characterize that dimension with a unique feature depending on the brightnesses of

both parts. Therefore, for each of the three stimulus sets, a linear combination of 3 components is found to be significantly correlated: Complex brightness = $\alpha.x_1 + \beta.x_2 + \gamma.x_3$, where x_1 is the Perceptual Spectral Centroid of the harmonic part, x_2 is that of the noise part and x_3 is the overall Perceptual Spectral Spread (see Appendix D). Both harmonic and noise parts were separated with the method and MATLAB code taken from Ellis [26]. The linear regressions of the obtained "complex brightness" with the second dimensions of the *motor* meta-category are shown in Figs. 4.1 to 4.3. However, no common combination was found to be correlated for every stimulus set. Tab. 4 shows the coefficients of this "Complex brightness" for each stimulus set.

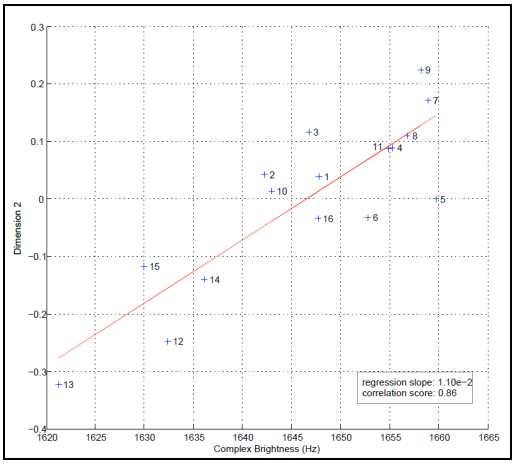

Fig. 4.1: Linear regression between dim. 2 and Complex Brightness, *motor* class / study A1.

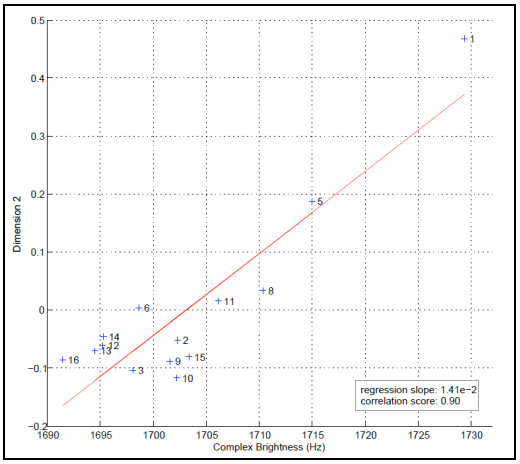

Fig. 4.2: Linear regression between dim. 2 and Complex Brightness, *motor* class / study A2.

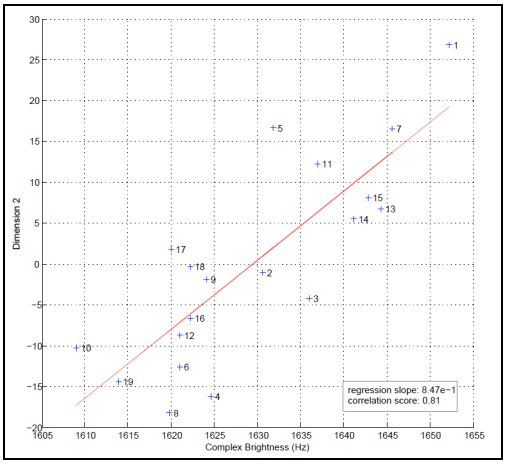

Fig. 4.3: Linear regression between dim. 2 and Complex Brightness, *motor* class / study B.

Study	α	β	γ
A1	+3.82e-3	1	-89.7
A2	+1.15e-2	1	-25.8
B	-1.08e-2	1	-63.0

Table 4: Coefficients of the linear combination defining Complex brightness for studies A1, A2 and B.

• Dimension 3

Study A1. This dimension seems to be well correlated with the **Perceptual Spectral Spread** – PSS (see Appendix D) calculated with logarithmic scales for both magnitude (level) and frequency. The linear regression of this feature with the third dimension of the perceptual space of this study is shown in Figure 5.

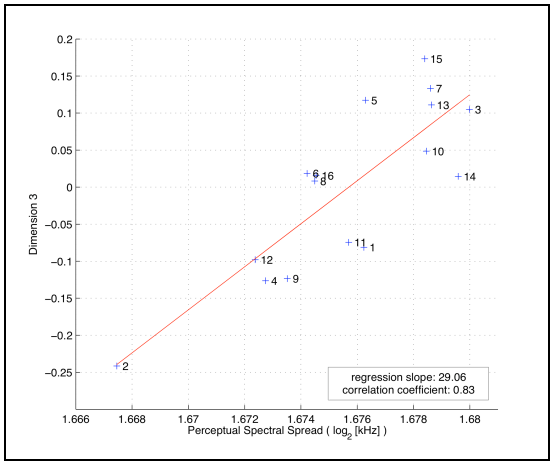

Fig. 5: Linear regression between dim. 3 and Perceptual Spectral Spread, *motor* class / study A1.

Study B. Unlike study A, the sounds were not initially loudness-equalized in study B. Quite logically, the last dimension of this MDS analysis result is found to be significantly correlated with **Loudness** (see Appendix D). The linear regression between Loudness and the third dimension of the study B perceptual space is shown in Figure 6.

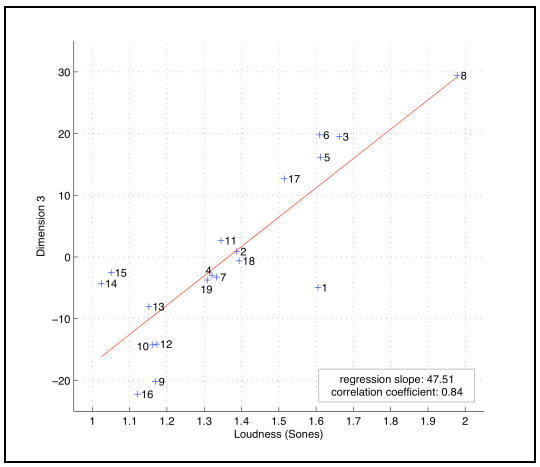

Fig. 6: Linear regression between dim. 3 and Loudness, *motor* class / study B.

Loudness is a perceptually strong characteristic that can easily prevent slight variations of other features from emerging. Moreover, the fact that no third perceptual dimension was obtained for stimulus set A2 can be related to the predominance of the noisy part, which can mask some variations of other features. On the contrary, when the sounds are loudness-equalized and when the harmonic part is not entirely masked by the noise, such as in stimulus set A1, a third perceptual dimension (PSS, Perceptual Spectral Spread) seems to emerge and matches that of the perceptual space of (pseudo-)harmonic instrument-like sounds (see Sec. 3.1.2 – Dimension 3). For these reasons, we were not able to unify this third dimension along the three corpora (A1, A2 and B).

3.1.3 *Instrument-like* category

This sound category corresponds to the stimulus set of study C. Its MDS analysis resulted in a 3-dimensional perceptual space presented in Sec. 1.2.3. According to the correlation scores in Tab. 5, those 3 dimensions are related to three different acoustic features presented below:

	Dimension 1	Dimension 2	Dimension 3
Roughness	**-0.93****	-0.06	0.37
Simple brightness (PSC)	0.04	**0.97****	0.05
PSS	0.05	-0.11	**-0.90****

Table 5: Correlations between acoustic features and dimensions of the *instrument-like* category / study C ($df = 20$, ** $p < 0.01$).

• Dimension 1: **Roughness**

Study C. This dimension seems to discriminate the monophonic from the polyphonic sounds. When listening to the sounds along this scale, one goes from perfectly harmonic tones to successively pseudo-harmonic tones (tones with inharmonicity relationships between their partials) and polyphonic sounds (with several tones). Consistently, roughness correlates significantly with this dimension (see Appendix D). The linear regression of roughness onto the first dimension is shown in Figure 7.

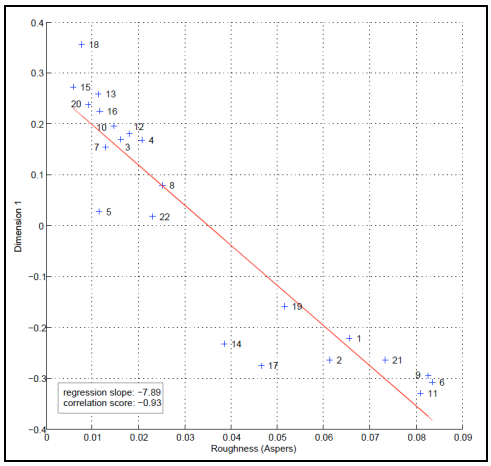

Fig. 7: Linear regression between dim. 1 and Roughness, *instrument-like* class / study C.

• Dimension 2: Perceptual Spectral Centroid (**PSC**) – *Simple brightness*

Study C. When listening to the sounds along this scale, the relation to the brightness of the sounds seems quiet obvious. This brightness is well quantified by the spectral centroid all the more when a perceptual model is used. Consistently, the Perceptual Spectral Centroid gives the best correlation score (see Appendix D). We call it *Simple brightness* because it can be formally seen as the degenerated form of the *Complex brightness* defined in the previous section, when harmonic and noise part of the signal are not separated. The PSC linear regression with the second dimension is shown in Figure 8.

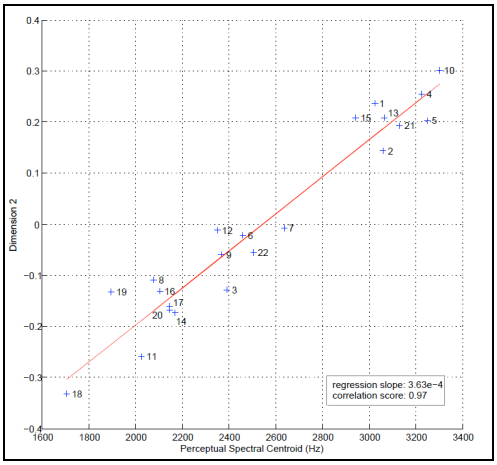

Fig. 8: Linear regression between dim. 2 and Perceptual Spectal Centroid, *instrument-like* class / study C.

• Dimension 3: Perceptual Spectral Spread (**PSS**)

Study C. This dimension is the one whose interpretation is the most difficult just by listening to the sounds along the scale. However, it could be associated with their "richness". It correlates quite well with the Perceptual Spectral Spread (see Appendix D). The linear regression of PSS onto the third dimension is shown in Figure 9.

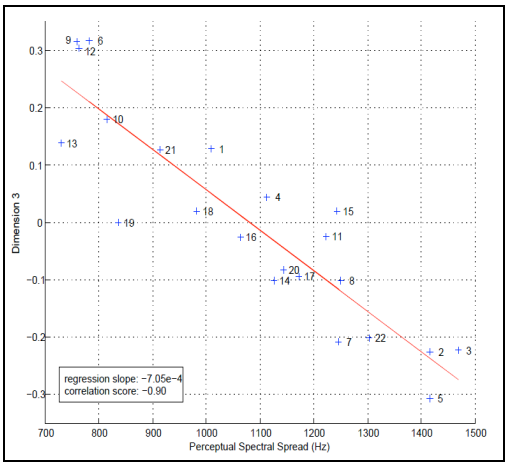

Fig. 9: Linear regression between dim. 3 and Perceptual Spectral Spread, *instrument-like* class / study C.

3.1.4 *Impact* category

This sound category corresponds to the stimulus set of study D. Its MDS analysis resulted in a 3-dimensional perceptual space presented in Sec. 1.4. According to the correlation scores in Tab. 6, those 3 dimensions are related to three different acoustic features presented below:

	Dimension 1	Dimension 2	Dimension 3
Simple brightness (PSC)	**-0.89****	0.05	0.08
Cleanness indicator	-0.18	**0.90****	0.24
RMS value	-0.27	0.18	**0.88****

Table 6: Correlations between acoustic features and dimensions of the *impact* category / study D ($df = 10$, ** $p < 0.01$).

• Dimension 1: Perceptual Spectral Centroid (**PSC**) – *Simple brightness*

Study D. The feature that best suits this dimension is the Perceptual Spectral Centroid (PSC) that includes a hearing model (see Appendix D). Indeed, this dimension describes the sounds' brightness. We call it *Simple brightness* for the same reasons presented in Sec. 3.1.2, regarding the second dimension of the *instrument-like* category. The linear regression between the PSC feature and the first perceptual dimension is shown in Figure 10. However, it is noticeable that there is a categorization phenomenon along this dimension, as the sounds labeled 9, 11 and 12 are much lower on that dimension than the other ones. This phenomenon comes from the MDS analysis results and is not only related to the tested features. Nonetheless, it tends to improve the correlation score.

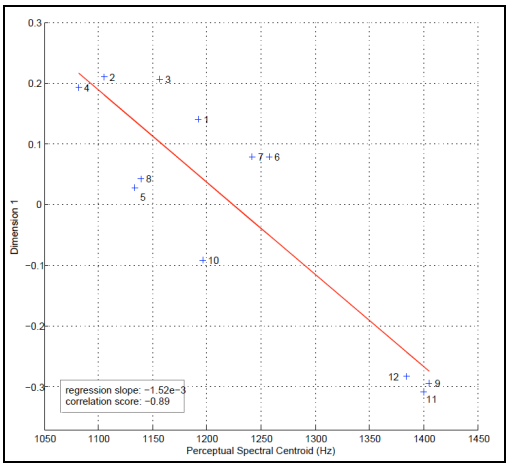

Fig. 10: Linear regression between dim. 1 and Perceptual Spectral Centroid, *impact* class / study D.

• Dimension 2: **Cleanness indicator**

Study D. It seems, when listening to the sounds along this scale, that this dimension is linked with the *cleanness* of the sounds. More precisely, it discriminates sounds containing only one impulse such as

the sounds numbered 1, 2 and 3, from those in which one or more impulses follow the main one (rattle, bounce...), such as the sounds numbered 10, 8 and 7. The acoustic feature (*Cleanness indicator*) that best suits this dimension is an estimator of the short-term loudness variability of the sounds (see Appendix D). This linear regression of the Cleanness indicator onto the second dimension is shown in Figure 11.

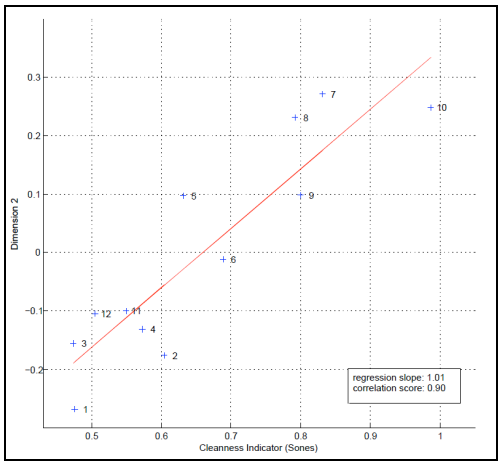

Fig. 11: Linear regression between dim. 2 and Cleanness indicator, *impact* class / study D.

• Dimension 3: **Sound level**

Study D. The RMS value is correlated with this dimension. Indeed, the dimension seems to be somehow related to pulse amplitude. The linear regression of this feature onto the perceptual dimension is shown in Figure 12

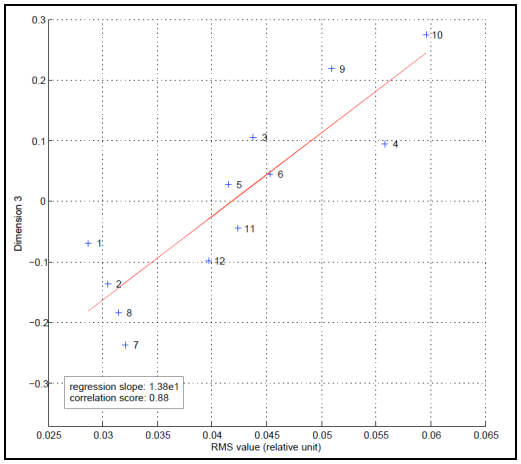

Fig. 12: Linear regression between dim. 3 and Sound level, *impact* class / study D.

3.1.5 Discussion

Looking for regularities and singularities among the 3 important categories of environmental sounds derived from the first part of this study, we finally identified:

• One feature, *Brightness*, that is preponderant for the description of all sound categories (i.e., 1 dimension of the 5 perceptual spaces)., This feature is actually a combination of different spectral envelope features: the perceptual spectral centroid of both harmonic and noise parts of the signal (PSC_h and PSC_n) – or perceptual spectral centroid of the whole signal (PSC) – and perceptual spectral spread (PSS). And no unique combination has been found to describe uniformly this dimension. So this feature still remains a generic notion of brightness and cannot be transformed into a real metric for quantifying this dimension.

• One or two features, in each category, that are related to specificities of the corresponding sounds:

– motor sound perception is largely characterized by the mixture of two highly discriminable parts, in terms of either energy or spectral content;

– instrument-like sounds present timbre features that have been found previously for musical sounds (essentially, roughness);

– an important part of the perceptual discriminability of impact sounds is related to a temporal behavior feature, describing the sounds' cleanness.

3.2 Categorical level: building an automatic classifier

Now that we have identified the inter-category specificities, we must address the development of a predictive tool able to automatically classify the sounds on the basis of a perceptually validated corpus. In other words, the aim here is to use the results presented in Sec. 3.1 as relevant cues in order to find a limited number of acoustic features that would be efficient for the implementation of an automatic perceptual classifier.

3.2.1 Specificities of the categories

Before considering the implementation of such a tool, it is essential to identify which features are used when listening to the sounds in order to discriminate the three categories. As partially concluded in Sec. 3.1.4, we can assume that:

• Impact sounds differ from the other ones in their temporal structure: they are quite short because they are damped, while the other sounds are as long as desired because they are sustained;

• Instrument-like sounds differ from the other ones in their spectral structure: their spectrum energy is usually localized in the middle frequencies and their spread is quite low, because they are harmonic sounds whose degree of spectral envelope decrease is high. To the contrary, the spectrum energy of the other sounds is localized in much lower frequencies with a much higher spread and a lower degree of

spectral envelope decrease.

Thus, it seems obvious that the cue that discriminates motor sounds from impact sounds, for instance, is very different from the one that discriminates motor sounds from instrument-like sounds. As a consequence, it is quite certain that a unique feature will not be enough to describe the categories, and it is more likely that we will have to use a pair of temporal and spectral features.

According to these preliminary obervations, a large set of temporal/spectral feature pairs could be used in order to discriminate the category to which a given sound belongs. Spectral and temporal features that seem to be good candidates for dealing with this problem are listed below. Their terminology and computing techniques are taken from Peeters [15]:

• Temporal features: Log-Attack-Time (LAT), Temporal Increase (TI), Temporal Decrease (TD), Temporal Centroid (TC), Effective Duration (ED), Energy Modulation Frequency (EMF) and Energy Modulation Amplitude (EMA);

• Spectral features: *mean* component of Spectral Centroid (SC), Spectral Spread (SSp), Spectral Skewness (SSk), Spectral Kurtosis (SK), Spectral Slope (SSl), Spectral Decrease (SD), Spectral RollOff (SR) and Spectal Variation (SV).

3.2.2 Classification modeling tool: the multinomial logistic regression

Now that we have identified the feature combinations that are likely to discriminate the three sound categories, we need a regression modeling tool able to predict the values of a qualitative and polytomous dependent variable Y (i.e., the sound category) by a combination of quantitative independent variables X_1, ... , X_k (i.e., acoustic features). This tool is the multinomial logistic regression (see Legendre et al. [22] and Woodcock [27]). In its basic definition, logistic regression is used to discriminate only two different attributes (or values) of a binary dependent variable Y (with values 0 and 1). With the probability notation $\pi(x) = P(Y = 1|X = x)$ of the event where the Y variable has the value 1, given the $x = (x_1, \ldots, x_k)$ value of the $X = (X_1, \ldots, X_k)$ set of variables, both event probabilities are related to each other by Eq. 1:

$$\pi(x) = P(Y = 1 \mid X = x) = 1 - P(Y = 0 \mid X = x) \quad (1)$$

A logistic regression tool models the $\pi(x)$ probability by a logistic function, formulated in Eq. 2. This function, which exhibits a sigmoid curve ("S-shaped" curve), is defined as the cumulative distribution function of a logistic probability distribution (similar to the normal distribution).

$$\pi(x) = \frac{1}{1 + e^{-u}} = \frac{e^{u}}{1 + e^{u}} \quad (2)$$

where u is a linear combination of the values of x: $u = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

Its inverse function, the "logit" function, corresponds to the natural logarithm of the odds' ratio in Eq. 3:

$$logit(\pi(x)) = \log(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \quad (3)$$

When the dependent variable Y corresponds to a polytomous nominal response (i.e., that has more than two different unordered values), the generalized logit models are used. In our case, the dependent variable Y corresponds to a 3-valued response: '0' for *impact*, '1' for *motor* and '2' for *instrument-like*. With the notation $\pi_i(x) = P(Y = i|X = x)$, the multinomial logistic regression consists in modeling the relationship between the set of independent variables $X = (X_1, \ldots, X_k)$ and the generalized logits, $\log(\pi_1(x)/\pi_0(x))$ and $\log(\pi_2(x)/\pi_0(x))$. The model assumes a linear relationship for each logit as in Eq. 4:

$$\begin{aligned} \log\frac{\pi_1(x)}{\pi_0(x)} &= \beta_{10} + \beta_{11} x_1 + \ldots + \beta_{1k} x_k \\ \log\frac{\pi_2(x)}{\pi_0(x)} &= \beta_{20} + \beta_{21} x_1 + \ldots + \beta_{2k} x_k \end{aligned} \quad (4)$$

The regression tool searches iteratively for the best-fitting solution (β_{ik} coefficients) using the Newton-Raphson method and maximum log-likelihood as a convergence criterion. The predicted probabilities are then given by Eqs 5.1 to 5.3:

$$\pi_1(x) = \frac{e^{u_1}}{1 + e^{u_1} + e^{u_2}} \quad (5.1)$$

$$\pi_2(x) = \frac{e^{u_2}}{1 + e^{u_1} + e^{u_2}} \quad (5.2)$$

$$\pi_0(x) = 1 - (\pi_1(x) + \pi_2(x)) \quad (5.3)$$

where $u_1 = \beta_{10} + \beta_{11} x_1 + \ldots + \beta_{1k} x_k$ and $u_2 = \beta_{20} + \beta_{21} x_1 + \ldots + \beta_{2k} x_k$

3.2.3 Model selection

This tool is applied to the perceptually validated sound corpus established at the end of Sec. 2, in order to predict the belonging of a sound to one of the 3 identified categories. This corpus is large enough (207 sounds) to make the results of such a procedure relevant. According to the set of acoustic features selected in Sec. 3.2.1, we can compute a classification model for each pair of spectral/temporal features. The best model's selection is made on the basis of their respective log-likelihoods. The log-likelihood LL is a statistical feature that corresponds to the sum of each natural logarithm of the predicted probability $\pi(x)$ that a sound belongs to its supposed category, as described in Eq. 6:

$$LL = \sum_x \log(\pi(x)) \quad (6)$$

However, the log-likelihood value depends on the number of elements within the stimulus set, and having the same value with stimulus sets of different size is not as relevant. A way to take this into account is to calculate a likelihood ratio that quantifies the gain in correct prediction of the model compared to the "intercept only" model, where only the β_0 constant coefficients are used[2]. The likelihood ratio feature LR is obtained with the relation defined in Eq. 7:

$$LR = -2 * (LL_n - LL) \quad (7)$$

where LL_n is the "intercept only" model log-likelihood. This statistical feature allows us to compare the effectiveness of each model (i.e., each feature pair) in predicting the category to which a given sound belongs. The higher the LL and LR values are, the more efficient the model is (see Legendre et al. [22]). The LR value for each feature pair is shown in Tab. 7, where we can see that the SSp/ED model seems to best suit the data.

	LAT	**TI**	**TD**	**TC**	**ED**	**EMF**	**EMA**
SC	167.1	252.7	333.5	373.0	380.5	289.8	101.8
SSp	187.0	257.1	354.6	399.2	**407.3**	317.5	122.7
SSk	106.1	180.2	314.5	373.1	385.7	235.5	35.7
SK	110.2	188.3	318.9	374.6	386.9	245.1	43.3
SSl	167.1	252.7	333.5	373.0	380.5	289.8	101.8
SD	108.7	187.0	295.9	355.9	369.9	215.8	22.1
SR	98.3	177.9	295.2	353.9	368.3	223.5	43.3
SV	115.6	171.1	306.8	360.9	376.1	224.6	74.0

Table 7: LR value for each spectral/temporal feature pair.
Spectral features are in rows and temporal features are in columns.

3.2.4 Model validation

In order to test the robustness of the selected model using SSp and ED features (see Sec. 3.2.3), a usual method consists of:

i/ re-estimating the model on a randomly selected reduced part of the stimulus set, 70% of it for instance (144 sounds with respect to the distribution in the 3 categories),

ii/ calculating the estimated probabilities on the remaining 30% (63 sounds),

[2] This means that the "intercept only" model will give the same probabilities whatever the data. In the present case, it will give a 99/207=0.48 probability of belonging to the *motor* category, a 49/207=0.24 probability of belonging to the *instrument-like* category and a 59/207=0.28 probability of belonging to the *impact* category.

iii/ evaluating the error percentage[3].

This procedure was performed 100 times with a different random selection of sounds each time. This method tests whether the effectiveness of the model prediction will hold when applied to other sounds than those used to estimate its coefficients.

When estimated on the whole 207-sound stimulus set, the best-fitting model makes 7 errors, which corresponds to an error percentage of 3.3%. Over the 100 times we performed the procedure explained above, we obtained the results presented in Tab. 8, calculated on the recall number (total number minus number of errors) of every remaining 30% selection of the stimulus set. One may observe that the mean recall percentage (95.9%) is rather high, not even much smaller than when obtained on the whole stimulus set (96.7%), which proves the model's adequacy for this dataset.

Minimum recall number	57
Minimum recall percentage	90.5%
Maximum recall number	63
Maximum recall percentage	100%
Recall number standard deviation	1.3
Mean recall number	60.4
Mean recall percentage	95.9%
Mean recall percentage interval	93.8% — 97.9%

Table 8: Results of the predicting tool based on SSp/ED features, after 100 runs of a 70%-learning/30%-predicting loop on the 207-sound perceptually extended corpus (Experiment 2).

3.2.5 Discussion

The selected model tested on a 207-sound stimulus set (augmented corpus established in Experiment 2, Sec. 2.2) gives significant stable results in terms of automatic classification with only around 4% mean error in the prediction, with only 2 predicting acoustic features. This is a rather encouraging result, even if this tool is built with only 3 main sound categories of quite different kinds (*motor*, *intrument-like* and *impact*). It could be extended to other categories in order to cover a larger scope of environmental sounds.

Other automatic classification methods exist that are much more complex and that use much more input information about the sounds. But considering the significant results of this relatively simple method,

[3] We consider as an error the case of a sound for which the probability of belonging to its supposed category is smaller than one of the two other probabilities. This means that if the model has to choose the category to which the sound belongs, it will choose a wrong one.

exploring these algorithms further is quite pointless. However, with more than 3 categories, these methods may outperform the one presented here and could therefore be useful for efficient automatic classification. From a larger point of view, other classification approaches also exist that are less time consuming with regard to the available data needed for performing them: they usually consist in defining sound classes, collecting training examples for each class, computing a large set of spectral and temporal features on sounds and letting a machine learning method pick features that are efficient in discriminating the classes. But, the main difference between this approach and the one proposed in the present paper relies on the fact that in the former, the classes are arbitrarily defined (or at least, are the result of a single expert's analysis), whereas in the present paper the classes are deduced from an experimental procedure, which is more time consuming but allows them to be considered as perceptually relevant. This is one of the original contributions of this study with regard to traditional methods based on *a priori* sound categories and powerfull learning techniques (*e.g.*, like the ones used in the Music Information Retrieval research[4])

3.3 Summary

We have built a 2-level description structure of environmental sounds that consists of:

- a categorical level that considers the different sound categories corresponding to particular sound production mechanisms,
- a continuous level that defines, within each of these categories, the perceptual space of the sounds allowing the represention of the perceptual dissimilarity between two sounds of the same kind.

This description is associated with automatic processing of acoustic features. When considering a new sound of one of these kinds, this processing allows: i) the identification of the sound category to which it belongs, with regard to the probabilities estimated by the logistic regression model, and ii) its correct placement along several perceptual dimensions.

Conclusion

This work originally aimed to extend timbre description principles, usually used for musical sounds, to environmental sounds and to apply them in a more systematic manner to this class of sounds. It is based on a first step of re-examination and comparison of four primary studies mainly dealing with industrial (car and machine) sounds. An inventory of their respective contexts, motivations, procedures and results gave us input data consisting of 5 coherent stimulus sets with their associated low-dimensional perceptual spaces. It also allowed us to intuit some regularities and singularities among the different kinds of sounds under consideration. Within the restricted scope of these 5 stimulus sets, a 2-part experimental approach

[4] http://www.ismir.net/

revealed 3 meta-categories (*motor*, *instrument-like* and *impact*) and precisely defined them in a larger scale by extending their respective contents. This categorical description structure is also coherent with the categories of product sounds that Özcan et al. [23] found. Finally, a modeling approach was designed to describe more precisely the intuited regularities and singularities of these 3 categories. This includes comparing the initial perceptual spaces by means of systematically correlated acoustic features, which can be summarized by two important facts:

- One feature is preponderant for the description of all sound categories, i.e., the *brightness* feature, usually based on spectral envelope features. Therefore, this perceptual feature appears to describe musical sounds as well as environmental sounds.
- One or two features, in each category, are related to a specificity of the corresponding sounds:
    - motor sound perception is largely characterized by the mixture of two highly discriminable parts, in terms of either energy or spectral content,
    - instrument-like sounds present timbre features, originally derived for the description of musical sounds,
    - an important part of the perceptual discriminability of impact sounds is related to a temporal behavior feature, describing the sounds' *cleanness*.

This modeling approach also includes the building of a predictive tool based on logistic regression able to classify automatically and rather efficiently (with only a 4% mean error) this meta-structure with regard to the 3 categories under consideration.

Note that contrary to musical timbre for which attack time is an important cue of the perceptual space, the studies revealed no temporal features corresponding to the two first categories. This may be mainly due to the quasi-stationary nature of these sounds. Nonetheless, a temporal parameter, associated with a spectral one, appeared to be fairly efficient in automatically discriminating impulsive environmental sounds (car door closing) from non-impulsive ones.

However, according to Özcan et al. [23], other major sound categories, such as liquid or cyclic sounds, exist and need a definition as well, and their main perceptual features must be investigated. Furthermore, they focused their study on domestic "product sounds", while we were more interested in industrial (cars and machine) sounds. Considering environmental sounds in a more general sense may again reveal other categories that would also need to be taken into consideration when building an overall environmental sound description structure, in terms of either definition or automatic description.

From an application point of view, the revelant acoustic features obtained for the three categories of sounds will allow us to conceive of perceptually relevant oganisation structures of large environmental sound collections and to propose retrieval systems using an intuitive query process by searching for sounds that are similar to a target sound in that kind of database. The search will be based on similarity

metrics computed from the acoustics features, stored with the sounds in the database as proposed by previous studies for musical sounds (Blum et al. [33]; Misdariis et al. [34]; Qi et al. [35]). In a larger perspective, these results should also contribute to the elaboration of a functional Computer-Aided Sound Design framework as they will help users to describe, associate, compare, share and finally manipulate sounds that can be considered as prototypes or initial ideas of concepts that the designer has in mind and tries to materialize in the framework of a specific project.

Acknowledgments

Some of these results come from the SampleOrchestrator project funded by the French Agence Nationale de la Recherche (ANR):
http://www.ircam.fr/306.html?&tx_ircamprojects_pi1%5BshowUid%5D=36&tx_ircamprojects_pi1%5BpType%5D=p&cHash=9859699b3d.

References

[1] Vanderveer N. J. (1979). Ecological acoustics: human perception of environmental sounds. *Unpublished doctoral dissertation*, Cornell University, pp. 16-17.

[2] Susini P., McAdams S. Winsberg S. (1997). Caractérisation perceptive des bruits de véhicules. *Proceedings of 4ème Congrès français d'acoustique*, Marseille, France.

[3] Susini, P., McAdams, S., Winsberg S. (1997). Perceptual characterisation of vehicules noises. *EEA Symposium: Psychoacoustic in Industry and Universities*. Eindhoven, The Netherlands.

[4] Susini, P., McAdams, S., Winsberg, S. (1999) A multidimensional technique for sound quality assessment. *Acta Acustica united with Acustica*, 85, 650-656.

[5] McAdams S., Susini P., Misdariis N., Winsberg S. (1998). Multidimensional characterisation of perceptual and preference judgements of vehicle and environmental noises. *Proceedings of Euronoise'98 conference*, Munich, Germany.

[6] Susini P., McAdams S., Winsberg S., Perry I., Vieillard S., Rodet X. (2004). Characterizing the sound quality of air-conditioning noise. *Applied Acoustics*, doi:10.1016/j.apacoust.2004.02.003.

[7] Lemaitre G., Susini P., Winsberg S., Letinturier B., McAdams, S. (2007). The sound quality of car horns: a psychoacoustic study of timbre. *Acta Acoustica united with Acustica*, 93, pp. 457-468.

[8] Lemaitre G., Susini P., Winsberg S., Letinturier B., McAdams S. (2009). The sound quality of car horns: Designing new representative sounds. *Acta Acustica united with Acustica* 95(2), pp. 356-372.

[9] Parizet E., Guyader E., Nosulenko V. (2006). Analysis of car door closing sound quality. *Applied Acoustics*, doi:10.1016/j.apacoust.2006.09.004.

[10] Grey J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, Vol. 61, No. 5.

[11] Krumhansl C. L. (1989). Why is musical timbre so hard to understand? in S. Nielzen & O. Olsson (Eds.), *Structure and Perception of Electroacoustic Sound and Music*, pp. 43-53, Elsevier (Excerpta Medica 846), Amsterdam.

[12] McAdams S., Winsberg S., Donnadieu S., De Soete G., Krimphoff J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological Research*, Vol. 58, pp. 177-192.

[13] Marozeau J., de Cheveigne A., McAdams S., Winsberg S. (2003). The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*., Vol. 114, No. 5, pp. 2946-2957.

[14] McAdams S. (1993). Recognition of auditory sound sources and events. in *Thinking in sound: the cognitive psychology of human audition*. S. McAdams and E. Bigand eds, Oxford University Press.

[15] Peeters G., (2004), A large set of audio features for sound description (similarity and classification). *CUIDADO project Ircam technical report*.
http://www.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf

[16] Slaney M. (1998). Auditory Toolbox, version 2. *Interval Research Corporation Technical Report* No. 1998-010. *http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/*

[17] Kruskal J. B., Wish M. (1978). Multidimensional scaling. *Sage University Paper series on Quantitative Applications in the Social Sciences* 07-011, Sage Publications, Beverly Hills, CA.

[18] Carroll J. D., Chang J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika*, 35:283-319.

[19] Winsberg S., De Soete G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika*, 58:315–30.

[20] Zwicker E. (1977). Procedure for calculating loudness of temporally variable sounds. *Journal of the Acoustical Society of America*, 62:675–81.

[21] McAdams S. (1994). La reconnaissance de sources et d'évènements sonores. *in Penser les sons*, chapitre 6, pages 157–213. Presses Universitaires de France. *(french translation of [14])*

[22] Legendre P., Legendre L. (1998), Numerical Ecology. *Development in environmental modelling.*

Elsevier, second english edition.

[23] Özcan E., van Egmond R. (2006). Memory for product sounds: the effect of sound and label type. *Acta Psychologica*, doi:10.1016/j.actpsy.2006.11.008.

[24] Rosch E. (1978). Principles of categorization. *in Cognition and Categorization*, chapitre 2, pages 27–48. Lawrence Erlbaum Associates.

[25] Smith, B. K. (1995). PsiExp: an environment for psychoacoustic experimentation using the IRCAM musical workstation. In *Society for music perception and cognition conference'95*. Univ. of Berkeley, CA.

[26] Ellis D. P. W. (2003). Sinewave and Sinusoid + Noise Analysis/Synthesis in Matlab. *http://labrosa.ee.columbia.edu/matlab/sinemodel/*

[27] Woodcock S. MATLAB econometrics toolbox. *http://www.sfu.ca/~swoodcoc/software/software.html*

[28] Zwicker E., Fastl H. (1990). Psychoacoustics: Facts and Models. *Springer Verlag*.

[29] Bogaards N., Roebel A., Rodet X, (2004). Sound Analysis and Processing with AudioSculpt 2. *Proceedings of International Computer Music Conference* (ICMC), Miami, USA , 2004.

[30] Slaney M. (1993). An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer Technical Report* 35.

[31] Rodet X. *The additive analysis-synthesis package*. Ircam tutorial, *http://recherche.ircam.fr/equipes/analyse-synthese/DOCUMENTATIONS/additive/index-e.html*

[32] Patterson R. D., Robinson K., Holdsworth J., McKeown D., Zhang C., Allerhand M. (1995). Complex sounds and auditory images. *Auditory Physiology and Perception*, Oxford, pp. 429 446.

[33] Blum, T., Keislar, D., Wheaton, J., and Wold, E. (1995). Audio Database with Content-based Retrieval. *Annu. Rev. Physiol.* 61, 457-476.

[34] Misdariis, N., Smith, B. K., Pressnitzer, D., Susini, P., and McAdams, S. (1998). Validation of a Multidimensional Distance Model for Perceptual Dissimilarities among Musical Timbres, *in ICA & ASA joint meeting (Journal of the Acoustical Society of America, Seattle, USA)*, p. 3005.

[35] Qi, H., Hartono, P., Suzuki, K., and Hashimoto, S. (2002). Sound database retrieved by sound. *Acoustical Science and Technology* 23, 292-300.

[36] Aures W. (1985). Der sensorische Wohlklang als Funktion psychoakustischer Empfindungsgrößssen. *Acustica*;58(5):282–90.

Appendix A.

MultiDimensional Scaling (MDS) analysis' principles

MDS models

MDS techniques represent the dissimilarity data by distances in a geometrical space. The simplest model represents the dissimilarity D_{ij} between two sounds i and j, averaged across the participants' ratings, by a Euclidean distance in a geometrical space with R dimensions (Eq. A1):

$$D_{ij} = \sqrt{\sum_{r=1}^{R} (x_{ir} - x_{jr})^2} \quad \text{(A1)}$$

where x_{ir} is the coordinate of sound i on the r^{th} dimension.

In this model, the space is rotationally invariant, which means that rotating its axes will not intrinsically change the space structure as long as they remain orthogonal.

The increasing sophistication of MDS techniques has led to a refinement of the initial model. This model, called INDSCAL (Individual Difference Scaling) (Caroll et al [18]), also considers the possibility that subjects weight the dimensions differently. It represents the dissimilarity D_{ij} between two sounds i and j, for each subject s by Eq. A2:

$$D_{ijs} = \sqrt{\sum_{r=1}^{R} w_{sr} \cdot (x_{ir} - x_{jr})^2} \quad \text{(A2)}$$

where w_{sr} is the weighting given by subject s to the dimension r.

Another refinement is proposed by the CLASCAL model (Latent Class Approach) (Winsberg et al. [19]). The dissimilarities are modeled as distances in an extended Euclidean space of R dimensions. Thus, the CLASCAL model postulates common dimensions shared by all stimuli, attributes particular to each stimulus (so-called *specificities*), and latent classes of subjects. Specificities account for the possibility that a sound may possess some unique feature that other sounds of the set do not share. Latent classes have different saliences or weightings for each of the common dimensions and for the whole set of specificities. For latent class t, the distance between two sounds i and j within the perceptual space is thus computed according to:

$$D_{ijt} = \sqrt{\sum_{r=1}^{R} w_{tr} \cdot (x_{ir} - x_{jr})^2 + v_t (s_i + s_j)} \quad \text{(A3)}$$

In Eq. A3, D_{ijt} is the distance between sound i and sound j, t is the index of the T latent classes, x_{ir} is the coordinate of sound i along the r^{th} dimension, w_{tr} is the weighting of dimension r for class t, R is the total number of dimensions, v_t is the weighting of the specificities for class t, and s_i is the specificity of sound i.

The class structure is latent: there is no a priori assumption concerning the latent class to which a given subject belongs. The CLASCAL analysis yields a spatial representation of the *N* stimuli on the *R* dimensions, with the specificity of each stimulus, the probability that each subject belongs to each latent class, and the weightings or saliences of each salient perceptual dimension for each class.

Moreover, in the INDSCAL and CLASCAL models, the presence of dimension weightings that differ between subjects or classes of subjects removes the rotational invariance of the obtained spaces, because the dimensions are fixed by the use of those weightings. As a consequence, it is assumed in both models that the dimensions of the space are perceptually meaningful.

Appendix B.

Complementary data and initial results related to the four primary studies.

Data related to study A1

	Dimension 1	Dimension 2	Dimension 3
RAPmv-A	**-0.81****	0.32	-0.33
CGg-ERB	0.35	**-0.7****	-0.14
Dec (266-2300)	-0.32	0.00	**-0.83****

Table B1: Correlation coefficients between the perceptual dimensions of study A1 and psychoacoustic descriptors (*df* =14, ** $p<0.01$).

• **RAPmv-A:** A-weighted harmonic-to-noise ratio. Both harmonic and noise parts were separated using additive analysis/synthesis (see Rodet [31], for more detail on the separation technique). The feature is the ratio of their levels expressed in dB(A).

• **CGg-ERB:** ERB Spectral centroid. The frequency dimension is represented in ERB-rate (distance in terms of Equivalent-Rectangular Bandwidth (ERB) filters; see Patterson et al. [32] and Slaney [30]).

• **Dec**: Harmonic spectral decrease. This feature is related to the shape of the spectral envelope computed from the harmonic components of the signal. In the present case, this feature is computed on the bandwidth of the spectrum, but represents the relative decrease in the envelope of the harmonic spectrum only between 266 Hz and 2300 Hz.

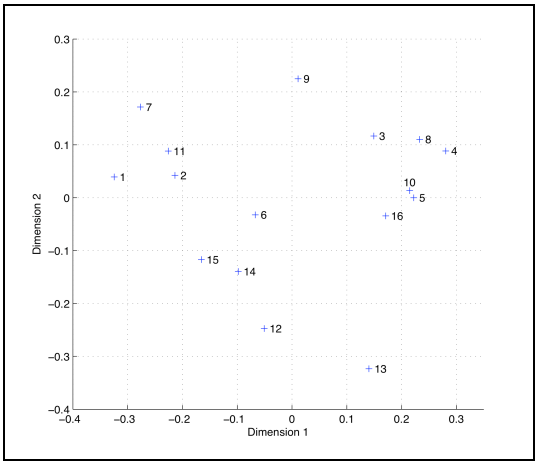

Fig. B1.1: Study A1 perceptual space projected onto dimensions 1 and 2.

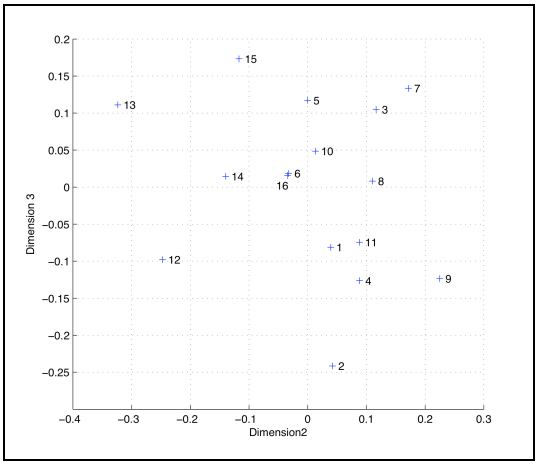

Fig. B1.2: Study A1 perceptual space projected onto dimensions 2 and 3.

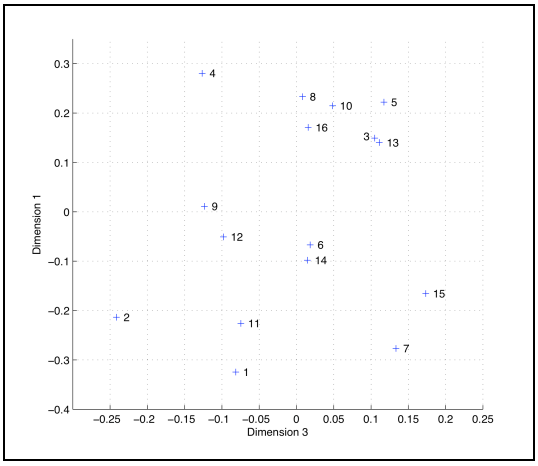

Fig. B1.3: Study A1 perceptual space projected onto dimensions 3 and 1.

Data related to study A2

	Dimension 1	Dimension 2
rad_2N/0.5N	**0.93****	-0.29
CGg-C	-0.51	**0.86****

Table B2: Correlation scores between the perceptual dimensions of study A2 and acoustic features (*df* =12, ** *p* <0.01).

• **rad_2N/0.5N:** 2N and 0.5N harmonics ratio, where N is deduced from the RPM value of engine rotation.

• **CGg-C:** Spectral centroid, with linear frequency using C-weighting.

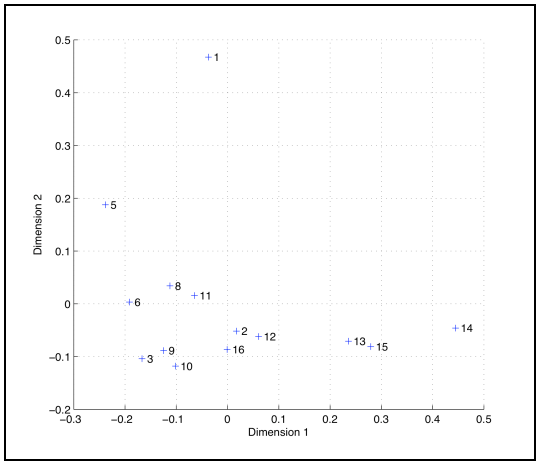

Fig. B2.1: Study A2 perceptual space projected onto dimensions 1 and 2.

Data related to study B

	Dimension 1	Dimension 2	Dimension 3
NHR-A	**-0.97****	0.11	-0.26
SC_n-B	-0.32	**0.73****	-0.15
N	0.26	0.04	**0.84****

Table B3: Correlation coefficients between the perceptual dimensions of study B and acoustic features (*df* =17, ** p <0.01).

• **NHR-A:** Feature corresponding to the relative balance of the harmonic (motor) and noise (air turbulence) components. The best correlation is obtained with the A-weighted version of this parameter.

• **SC_n-B:** B-weighted spectral centroid of the noise component. For this dimension, the emergence of a spectral pitch led us to consider the spectral centroid (SC). More precisely, we compute the SC of each of the two parts of the sound: the noise component (SC_n) and the harmonic component (SC_h). The best correlation with Dimension 2 is obtained for SC_n using B-weighting.

• **N:** Loudness. Indeed, even though the selected sounds are in the same range of loudness, they were not equalized in loudness.

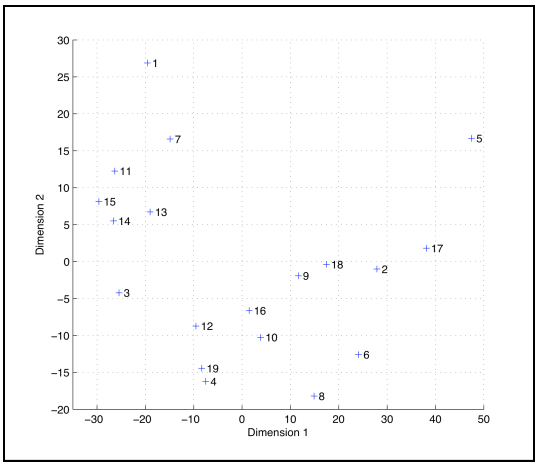

Fig. B3.1: Study B perceptual space projected onto dimensions 1 and 2.

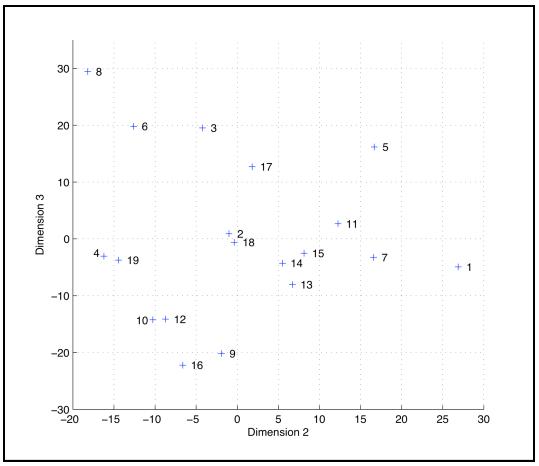

Fig. B3.2: Study B perceptual space projected onto dimensions 2 and 3.

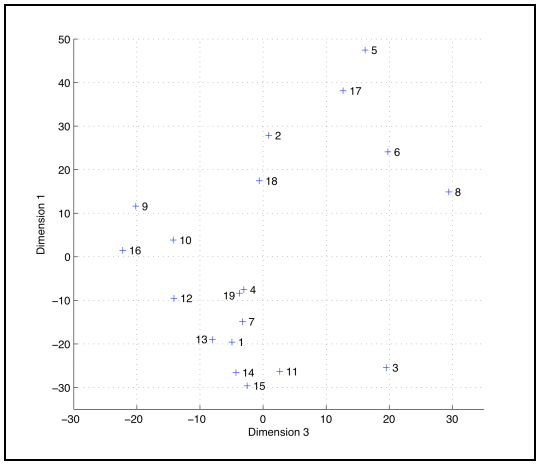

Fig. B3.3: Study B perceptual space projected onto dimensions 3 and 1.

Data related to study C

	Dimension 1	Dimension 2	Dimension 3
Roughness	**-0.9****	-0.1	0.3
Spectral centroid	0.0	**0.9****	0.1
Spectral deviation	0.3	-0.4	**-0.8****

Table B4: Correlation coefficients between the perceptual dimensions of study C and the best-correlated psychoacoustic descriptors (df=20, ** p<0.01).

• **Roughness:** Feature modeled by the amplitude modulation rate of the temporal envelope (expressed in asper) and related to the sensation of auditory roughness.

• **Spectral centroid:** Feature describing the spectral distribution of the energy of the sound, computed from a frequency decomposition on the ERB scale (Marozeau et al. [13]). It has been identified as corresponding to the sensation of "brightness".

• **Spectral deviation:** Feature related to the fine structure of the spectral envelope. It is computed based on the smoothness of the outputs of the filter-bank (Marozeau et al. [13])

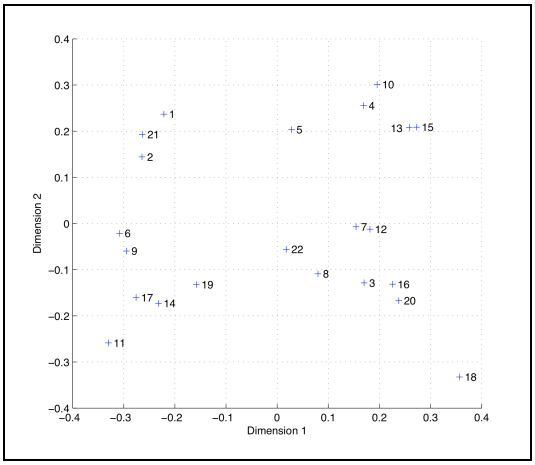

Fig. B4.1: Study C perceptual space projected onto dimensions 1 and 2.

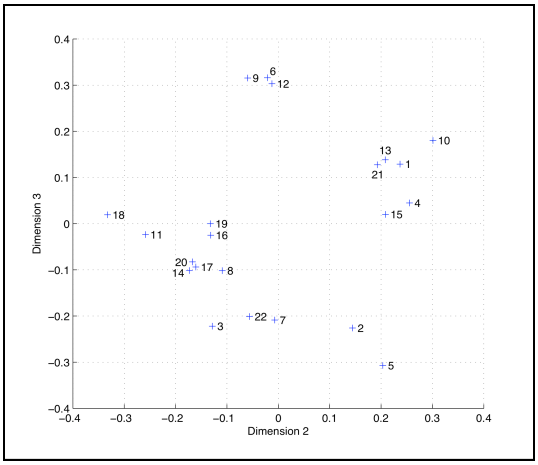

Fig. B4.2: Study C perceptual space projected onto dimensions 2 and 3.

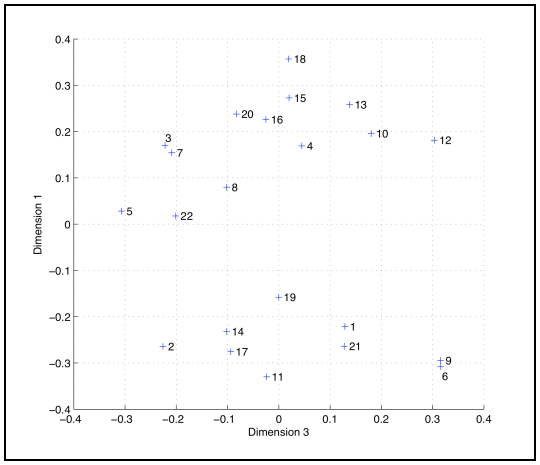

Fig. B4.3: Study C perceptual space projected onto dimensions 3 and 1.

Data related to study D

	Dimension 1	Dimension 2	Dimension 3
Sharpness	**-0.90**		
Spectral centroid	**-0.93**		
Cleanness indicator		**0.87**	
……			……

Table B5: Correlation coefficients between the perceptual dimensions of study D and acoustic features (*df* =10).

• **Spectral centroid**: Feature describing the spectral distribution of the energy of the sound.

• **Sharpness** Feature defined by Aures [36], similar to spectral centroid with perceptual modeling.

• **Cleanness indicator:** Indicator that is derived from the temporal loudness calculation according to Zwicker's model [20]. The algorithm takes into account temporal integration and temporal masking. The proposed indicator is based on the temporal evolution of the curve.

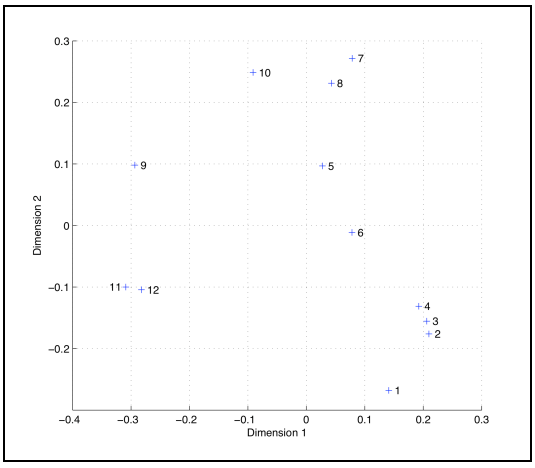

Fig. B5.1: Study D perceptual space projected onto dimensions 1 and 2.

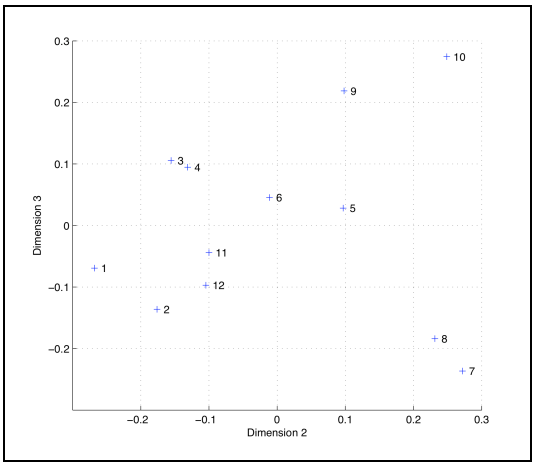

Fig. B5.2: Study D perceptual space projected onto dimensions 2 and 3.

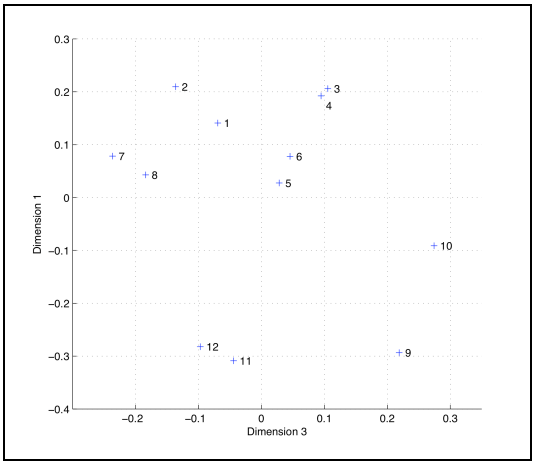

Fig. B5.3: Study D perceptual space projected onto dimensions 3 and 1.

Appendix C.

Illustration of the experimental graphical user interfaces used in Experiments 1 and 2.

Screenshot of Experiment 1 - GUI

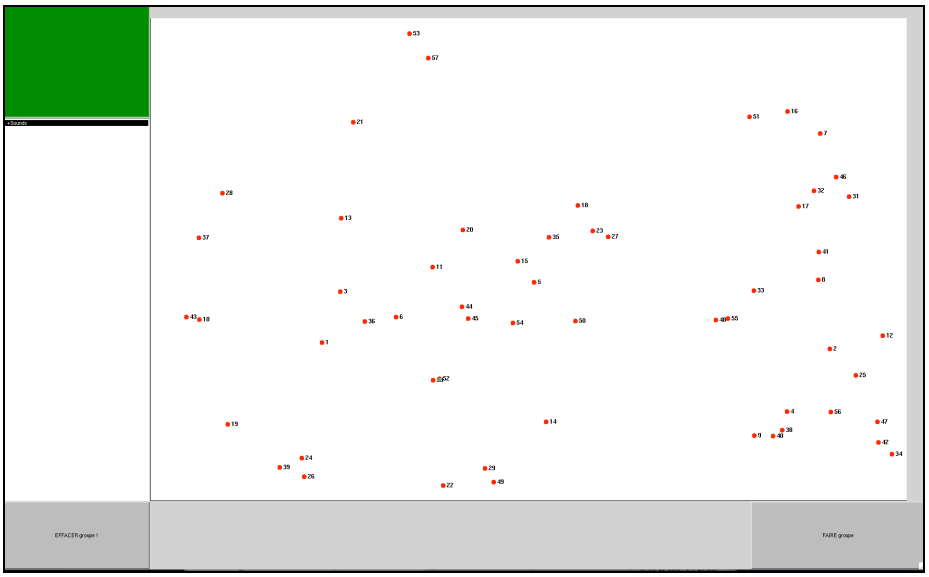

Fig. C1: Experiment 1 - GUI for free-sorting task.

Screenshot of Experiment 2 - GUI

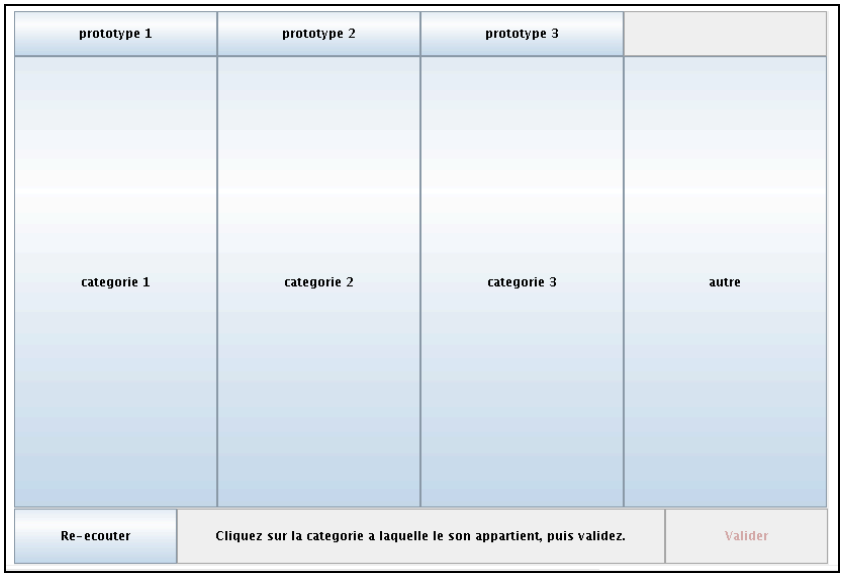

Fig. C2: Experiment 2 - GUI for the forced choice sorting task.

Appendix D.

Details of acoustic features calculation

RMS value

The estimation of the RMS (Root-Mean-Square) value of the signal is frame-based and is calculated every 60 ms with a Blackman window. The feature is the mean value over time.

Loudness

Loudness is the intensive attribute of human hearing. It thus describes the subjective aspect of the intensity of a signal by considering masking effects that occur over the whole spectrum and the filtering steps of the hearing path. The loudness model used is the ISO 532-B model from Zwicker et al. [28].

Harmonic emergence feature

This feature is a Harmonic-to-Noise ratio, designed to convey the relative amounts of harmonic (or pseudo-harmonic) energy and noise energy in the signal. It is based on the *Pm2* partial extraction method (see Bogaards al. [29]). Once both harmonic and noise parts of the signal are extracted, the feature simply consists of the ratio of their respective loudnesses N_h and N_n as formalized in Eq. D1:

$$HNR = N_h / N_n \tag{D1}$$

Spectral centroid

The spectral centroid is a weighted mean frequency of the spectrum of the signal. The calculation of this feature can be more or less complex. Its definition is quite similar to Zwicker et al.'s [28] *sharpness* feature. It uses a gammatone filterbank (from Auditory Toolbox, Slaney [16]) that is based on the ERB-rate scale *z* (see Marozeau et al. [13] for more details). The resulting feature is the Perceptual Spectral Centroid as defined in Eq. D2:

$$PSC = \sum_z f_z \cdot N_z \Big/ \sum_z N_z \tag{D2}$$

where N_z is the specific loudness in each channel (obtained by each gammatone filter) and f_z is the corresponding center frequency.

Spectral spread

The spectral spread describes how the spectrum is spread around its mean value, i.e. the spectral centroid

defined above. The associated perceptual feature uses the same perceptual modeling as the *PSC* feature, thus giving the Perceptual Spectral Spread *PSS*, as defined in Eq. D3

$$PSS = \sum_z (f_z - PSC)^2 \cdot N_z \Big/ \sum_z N_z \qquad \text{(D3)}$$

Complex Brightness

This feature estimates the *brightness* sensation of a sound that combines a noisy and a harmonic part. It simply corresponds to the linear combination of the *PSC* values of both noisy and harmonic parts (respectively PSC_h and PSC_n) and the *PSS* value of the whole signal, as defined in Eq. D4:

$$Complex\ \ brightness = \alpha \cdot PSC_h + \beta \cdot PSC_n + \gamma \cdot PSS \qquad \text{(D4)}$$

where α, β and γ are linear coefficients.

Roughness

Roughness is a feature that quantifies the perceived modulation or graininess of a sound. When inharmonicity is strong, amplitude modulations can generate beating in some cases. When the beating becomes fast enough so that the modulations are no longer discriminated by the human ear, they seem to give a *rough* aspect to the sound. This roughness feature (also defined in Grey et al. [10]) mainly consists in estimating a modulation index at the output of every auditory filter, which is called the *partial roughness*. The overall roughness is the sum of all the partial roughnesses. From each auditory filter output, the modulation frequency $f_{\text{mod}i}$ and the modulation depth m_i are estimated with a temporal envelope calculation. The partial roughness is proportional to the product of the modulation frequency and the depth $f_{\text{mod}i} \cdot m_i$. The roughness R is then calculated as the sum of the R_i, as mentioned in Eq. D5:

$$R_i = K \cdot f_{\text{mod}\,i} \cdot m_i \quad ; \quad R = \sum_i R_i \qquad \text{, where K is the proportionality coefficient.} \qquad \text{(D5)}$$

Cleanness indicator

This feature represents the short-term variations in the loudness of the signal. These variations, which usually occur between 20 and 100 Hz, are slow enough to be heard as a temporal phenomenon, but they are too fast to be heard as separate sound events (*e.g.,* bounces, rattles, etc.). The feature corresponds to the amplitude of the spectrum of the instantaneous loudness $N(t)$, which is estimated every 3.3 msec., within this frequency band (see Eq. D6).

$$Cleanness\ \ indicator = \sum_{20\text{–}100\,Hz} \left| FFT_{256}(N(t)) \right| \qquad \text{(D6)}$$

where FFT_{256} is the 256-point Fast Fourier Transform.

TABLE OF CONTENTS