# Towards computational morphological description of sound

by

**Julien Ricard**

Submitted in partial fulfilment of the requirements for
the degree of Diploma of Advanced Studies
Doctorate in Computer Science and Digital Communication
Department of Technology

Tutor: Dr. Xavier Serra

Universitat Pompeu Fabra

Barcelona, September 2004

# Abstract

Research on audio content description deals with limited types of sounds. Most of the work done in this area is applied to automatic transcription of traditional western music, i.e. the conversion of audio into the traditional musical notation pitch/duration/loudness/source or the recognition of the origin of specific sounds (speech, music, applause...) for indexing or retrieval purpose. In that context, electronic sounds, noises or sounds that have no identifiable origin, which are used a lot in contemporary music and sound post production for video or cinema, can hardly be handled. In this document, we propose an alternative representation, inspired by Pierre Schaeffer's work on *sound objects*, based on a limited number of perceptual criteria that can be applied to any type of sound. More specifically, we describe our first attempt to automatically characterize some of these criteria, called morphological criteria, as well as an evaluation of the usability of the resulting representation in the context of sound retrieval. Conclusions drawn from these experiments to improve and complete the system, as well as a the description of potential applications, are presented as future work to be done in the final thesis.

# Acknowledgments

I sincerely thank Xavier Serra for giving me the opportunity to work at the Music Technology Group and Perfecto Herrera for his constant support and guidance. I would also like to thank Nicolas Wack, Markus Koppenberger and the CLAM team for their technical help, as well as Fabien Gouyon for his useful comments on this work.

*I am sure that the time will come when the composer, after he has graphically realized his score, will see this score automatically put on a machine which will faithfully transmit the musical content to the listener. As frequencies and new rhythms will have to be indicated on the score, our actual notation will be inadequate. The new notation will probably be seismographic. And here it is curious to note that at the beginning of two eras, the Mediaeval primitive and our own primitive era (for we are at a new primitive stage in music today) we are faced with an identical problem: the problem of finding graphic symbols for the transposition of the composer's thought into sound. At a distance of more than a thousand years we have this analogy: our still primitive electrical instruments find it necessary to abandon staff notation and to use a kind of seismographic writing much like the early ideographic writing originally used for the voice before the development of staff notation. Formerly the curves of the musical line indicated the melodic fluctuations of the voice, today the machine-instrument requires precise design indications.*

The Liberation of Sound, Edgar Varese (1936)

# Table of Contents

# 1 Introduction

In this chapter we introduce the general context in which this work has been carried out and present our objectives as well as the outline of this document.

## 1.1 Context

Since the beginning of the 20th century and the first experiments on sound synthesis and transformation, electronic sounds have been an increasing part of our sound environment. At this time, sound processing machines were very large, expensive and allowed only a limited number of sound parameters to be controlled, often in a non-intuitive way, so that they were owned only by some laboratories or a few musicians[1]. Digital processing has made this technology accessible to a much larger range of people and has opened new perspectives in the composition of music, soundtracks or any other audio material. Electronic music is now accepted as a fully-fledged musical genre and has met a great commercial success, and synthesised or processed sounds are widely used in video or cinema soundtracks to create specific sound effects.

In addition to easing the generation of new sounds, digital technology allows storing a larger and larger amount of audio information in personal computers, audio libraries or archives, or on the web. Many tools have been developed in order to manage this data, mostly based on the transformation, analysis or retrieval of some representations manually set or automatically extracted from raw digital data. Most of the attention in this area is paid to music, and more particularly to the design of automatic music transcription systems (e.g. [Martin96] or [Klap04]) aiming at converting an audio signal into a symbolic representation, typically a score, and at extracting high level musical features from it. In the case of traditional western music, the representation consists of well-defined elements, the notes, described by a starting time, a pitch, a duration, a loudness and the instrument by which it is played. High-level analysis, such as rhythmic (see a review in [Gouyon03]), melodic (see a review in [Gomez03a]) or musical structure (e.g. [Dan02] or [Ong04]) description are derived from this representation or directly from the audio signal. Potential application of automatic description systems are numerous. Melody or rhythm, for instance, can be used as search criteria for retrieving a song which one has forgotten the name of or as parameters to be transformed to transpose or modify the tempo of a song. An important effort in the area of audio content description is also devoted to sound source identification, especially in the context of sound retrieval (see a review of

1 See http://www.obsolete.com/120_years/ for an history of electronic musical instruments.

musical instrument classification in [Herrera03] and of sound effect classification in [Zhang99]). In audio libraries, sounds are manually labelled according to their origin (e.g. 'dog bark') or to the musical notation when appropriate (e.g. 'C4, ff, piano') in order to provide keywords for the search. In large libraries, this task is very time-consuming, and performing it automatically would be of great interest. Electronic sounds, which exhibit a range of properties much wider than traditional musical sounds (typically harmonic sounds with constant pitch), and that have no identifiable source (except when they aim at reproducing natural sounds) are somewhat left apart by the audio content description community.

In this research work we investigate a representation, based on perceptual criteria, that could allow handling electronic sounds, as well as all non-traditional musical sounds and sounds having no identifiable origin, in audio content analysis, processing or retrieval tasks. This representation, called *typo-morphology* (see 2.4), was first defined by Pierre Schaeffer, a French musician, researcher and writer, in the late 60's, to provide a basis to the composition and the analysis of a musical genre making use of any king of sounds (noises, environmental sounds, loops...).

This work was initiated in the context of the CUIDADO[2] project on audio content description and has been carried on for the AudioClas[3] project on automatic classification of sound effects.

A list of publications related to this work is given in Appendix 3.

## 1.2 Objectives and document outline

The main objectives of this work are to show that a perceptual representation would be of great interest to handle non-traditional musical sounds and sounds having no identifiable origin - which will be referred to in this document as *abstract sounds* - in applications based on audio content, and to investigate what could be such a representation and how it could be automatically extracted.

After a general introduction to auditory perception, we give in the second chapter an overview of the research done on perceptual dimensions of sound and show that Schaeffer's description criteria, called *morphological criteria*, provides a good basis for a general perceptual description scheme.

We review in the third chapter the computational models proposed by the research community to mimic some features of human auditory perception. We also show that

2 http://www.ircam.fr/produits/technologies/multimedia/cuidado-e.html
3 www.audioclas.org

the only existing standard for audio description, the MPEG-7 audio, lacks general perceptual description criteria and that it should be completed to handle all kind of sounds.

In the fourth chapter, we describe a system, based on some models reviewed in the second chapter, that automatically extracts a simplified perceptual representation which each dimension, inspired by Schaeffer morphological criteria, is characterized either by a class textual label (e.g. dynamic profile = 'impulsive') or by a numerical value (e.g. roughness = 0.5).

The goal of the fifth chapter is to evaluate the usability of such a representation in the context of sound retrieval. More specifically, we present the results of a questionnaire showing that morphological criteria would provide useful keywords to retrieve abstract sounds in large sound effects libraries.

In the fourth chapter we describe some potential applications of computational morphological description and show that it is of great interest for many applications based on audio content, in which abstract sounds can currently hardly be handled.

Finally, we summarize and discuss the work described in this document and suggest guidelines for future research.

# 2 Describing sounds

The goal of this chapter is to provide an introduction to auditory perception and to review the perceptual dimensions of sounds identified by the research community.

## 2.1 Introduction to auditory perception and cognition

Sound can be heard or interpreted in different ways. When hearing speech, for instance, a listener generally focuses on what is being said in order to grasp the meaning of a message. Sometimes, one can also try to identify some characteristics of the speaker (e.g. sex, age...) or the speaker itself. A music lover would rather appreciate the particular timbre of his favorite singer's voice and the emotions he feels about it. All these processes -understanding, source identification or sound characterization- lead to different mental representations of sound. From the acoustical signal reaching the ear to these mental representations, one can identify two levels of analysis performed by our auditory system: a perceptual level and a cognitive level. Although the boundary between perception and cognition is quite debatable[4], the distinction proposed by the mainstream view, often referred to as the psychoacoustical auditory perception theory, is useful for our research.

Perception consists in the extraction of 'useful information' from our environment. It is the result of an evolution process that "designed" an auditory system specific to our needs. These needs are different for each species so that all animals do not perceive the world the same way. This 'useful information' is extracted by the human peripheral auditory system, shown in figure 2.1, and coded as an electrical signal transmitted by the auditory nerve to higher auditory structures. According to psychoacousticians, this signal leads to intermediate representations of sound, also referred to as 'mid-level representations', which are further organized and interpreted by cognitive processes. Cognition, from the Latin *co-gnoscere,* literally "to come to know", is defined by the encyclopaedia Britannica[5] as "the act or process of knowing including both awareness and judgment". While perception is sensory-specific and does not involve memory, cognition is amodal and involves memory, or knowledge, to interpret the information extracted at the perceptual level. According to that view, sound qualities such as pitch or timbre, which does not, a priori, require some knowledge, are perceptual criteria. On the other hand, speech understanding and speaker identification, which require, respectively, to know the spoken language and

4 A discussion on the difference between perception and cognition, started in April 2004, can be found in the archives of the Auditory mailing list (http://www.auditory.org ).

5 http://www.britannica.com

to have in memory some representation of the speaker, involve cognitive processes [6].

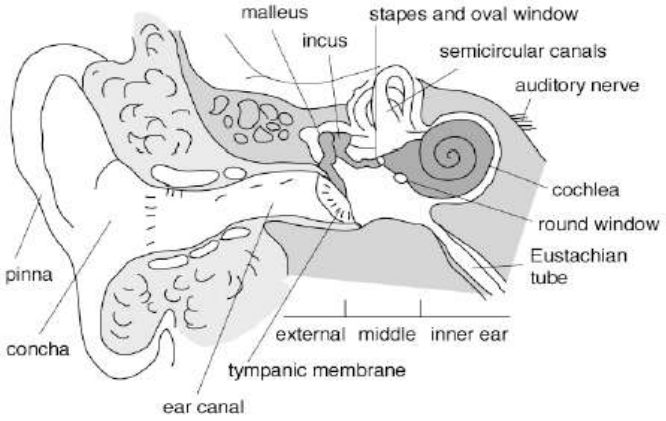



*Figure 2.1: Human peripheral auditory system. After M. Karjalainen's lecture material. Reprinted with the permission of the author.*

Our aim here is not to debate about the definitions or the difference between perception and cognition, but rather to clarify what we mean in this document by 'perceptual' criteria.

Introductions on auditory perception and cognition are given in [Handel89], [Yost00] and [McAd93] (auditory cognition only).

6 The processes described here, involving intermediate mental representations and their interpretation by cognitive processes, are the basis of the psychoacoustical perception theory. An alternative theory, based on Gibson's direct perception [Gibson66], and referred to as ecological perception, is also paid a lot of attention by the research community. This theory claims that no intermediate representations or memory are needed and that the brain directly recognize in the stimulus some features that specifies the sound. More specifically, some neural structures would 'resonate' to patterns specific to the environmental object itself (structural invariant, for instance a piece of wood) and to the action which makes it generating sound (transformational invariant, for instance the action of breaking). Although claiming that no representation or memory are involved (but aren't invariants some kind of representation, and isn't the resonance of neural structures some kind of memory process?) is quite a radical point of view, ecological perception have been a source of inspiration for many researchers looking for some computational way to extract invariants [Casey98] [Rocch03].
A review of auditory perception theories with a specific focus on ecological perception is given in [Macph95].

## 2.2 So what do we hear in a sound?

We saw that sound can be described according to two levels of analysis, corresponding to different levels of abstraction of the sound. In the following section we further review the different mental representations that constitute our conscious experience of the sound.

### 2.2.1 Schaeffer's four listening modes

In his 'Traité des objets musicaux' (Treatise on musical objects), [Schaef66] identifies four listening modes, each associated to a specific listening intention and a specific representation :

- *Ecouter* refers to the identification of the sound-producing event through the sound. In that case the sound is seen as an index of this event.
- *Comprendre* (to understand) refers to the identification of a message transmitted by the sound. This listening mode is well illustrated by speech, in which the sound is only the 'vehicle' of words, which themselves carry a meaning.
- *Entendre* refers to listening to intrinsic properties of the sound. When choosing an instrument, for instance, a musician will select the one that sounds the best for him, according to his own appreciation of its timbre.
- Ouir is the lowest level in our auditory perception. It refers to to the perception of the raw-sound data with no intention of interpreting or qualifying it. It is a kind of 'passive listening'.

Ouir and ecouter, according to Schaeffer, are spontaneous and universal ways to listen to sound and are referred to as natural listening. Comprendre and ecouter both require some references to codes, language in the case of speech and some description criteria in the case of the musician choosing an instrument. Since these codes must be learned and are not universal, Schaeffer qualified both these listening modes as cultural listening.

A slightly different dichotomy in the listening attitudes is described in [Gaver93], in the context of ecological perception. He defined the spontaneous focus on attributes related to the identification of the sound-producing event as 'everyday listening' (Schaeffer's 'écouter'), as opposed to 'musical listening', which defines, according to him, a focus on traditional perceptual attributes such as pitch or timbre.

### 2.2.2 How do we talk about sounds?

Schaeffer's identification of different listening modes was base in his own experience. A more systematic way to investigate how people experience sounds is to study how they talk about them. A major work on the subject, though limited to musical sounds, is described in [Faure00]. Among other experiments, she asked subjects a free verbal description of dissimilarity in pairs of sounds. Stimulus consisted mostly of synthesised sounds imitating traditional western musical instruments and a few hybrid sounds synthesised by morphing two known instruments. All sounds were equalised in duration, pitch and loudness.

Interestingly, for 60% of the pairs of sounds, dissimilarity judgments contains terms related to the origin of the sounds (material, action, name of instruments...), though it was not specified in the instructions. This confirms the natural tendency to 'listen to the source'. However, in that case, since the instruction was to talk about the sounds, we think that the source was named not for itself but for the set of perceptual attributes it refers to.

In 98% of the pairs[7], similarity judgments actually contained words related to the sound itself. Most of the terms (80%) referred to the temporal evolution of energy concerning, in equal proportion, the beginning ('slow attack', 'impact', 'fast'...), the middle ('stagnation', 'bouncing'...) and the end of the sound ('cut', 'muffled'...). Very often (73%), subjects described the sounds according to criteria related to the five senses ('high/low pitched', 'nasal', 'bright', 'rough', 'acid'...). For half of the pairs, they talked about some 'shape' of the sound ('large', 'sharped', 'compact'...). Other verbalisations were related to emotion or a judgment based on emotion ('pleasant', 'boring', 'nice'...), and sometimes the sounds were given a name ('note', 'crackling'...) or were even imitated.

As one could expect, subjects found more difficult to talk about the hybrid sounds, not perceived as being generated by a known source, and tended to talk about the sounds themselves. However, since these sounds were synthesised by morphing two instruments, some identity of how they were built remained and subjects sometimes described them as the juxtaposition of two sounds or could even recognize the original sounds.

The study concerns traditional musical sounds and must not be generalised to all sounds. However, a similar study on free verbal description of environmental sounds [Vander79] also observed this tendency for people to talk about the source of the sound..

7 Several descriptions could be used for each pair, so that the percentages do not add up to 100.

## 2.3 Source identification vs. sound description

We saw that listeners spontaneously describe sounds by referring to the sound-producing event. It is the primary function of auditory perception, and of perception in general, to provide us with some information about our environment in order to locate food, prevent from a danger or communicate... As we already mentioned, our auditory system has probably evolved to optimally perform these tasks and what we call perceptual attributes would then be the dimensions of an optimal space for recognizing events from our environment. However, our purpose is not to find some mapping between environmental sources and a perceptual space, i.e. to identify the source, but rather to investigate this space itself, i.e. to actually describe sound. As pointed out by Schafer, quoting Schaeffer, "The sound object must not be confused with the sounding body by which it is produced, for one sounding body may supply a great variety of objects whose disparity cannot be reconciled by their common origin." ([Schafer77], p130).

## 2.4 Describing sounds

We saw in Faure's experiment that talking about sound attributes is difficult, so that people often use analogies to others senses (bright, large...), for which more specific vocabulary exist. Very few attempts to define a general perceptual space for sound have been done. Most studies deal with sounds from specific origins (typically musical instruments or environmental sounds) and then only investigate a limited perceptual space (for instance traditional musical sounds are all harmonic while environmental sounds are mostly noisy). Our aim here is to review the sounds attributes that have been used for perceptual description of sound.

### 2.4.1 Traditional musical notation

A well known sound representation is the western musical notation, in which each sound is described by a pitch (horizontal position on the staff), a loudness (a symbol corresponding to a value on a discrete loudness scale), a duration and the instrument by which it should be played. The first note of figure 2.2 represents a C4 (261 Hz) with duration twice shorter than the second note and played mezzo-forte (moderately loud) by a piano.


*Figure 2.2: Traditional western music notation.*

This representation does not really aim at describing a sound and is rather a kind of recipe to make it. Although it contains important perceptual cues about the aimed sound (pitch, loudness and duration), it is still very limited: it only applies to harmonic sounds with constant pitch, and most of the sound attributes, often referred to as the 'timbre' of the sound, are only indirectly described by reference to a musical instrument.

## 2.4.2 Perceptual attributes of sound

### Subjective duration, loudness and pitch

An important feature of our auditory system is to convert physical scales to scales more appropriate to our understanding of the environment. The perceptual counterparts of duration, intensity and frequency are referred to as, respectively, subjective duration, loudness and pitch.

Duration is defined as an objective, physical measure of time in seconds, minutes or hours. However, the subjective perception we have of it seems to be quite dependent of what happens during this time. Experiments on subjective duration, using sequences of silences and bursts are described in [Zwicker90].

Loudness is defined by the American National Standards Institute (ANSI[8], 1973) as the “attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud". Loudness perception is highly level and frequency dependent. It has been studied by asking listeners to match the loudness of tones at various frequency to that of a reference tone, a 1000 Hz sinusoid, for various physical level [Fletcher33] [Zwicker90]. All tones perceived as being as loud as the reference tone at 40dB SPL have a loudness level of 40 phons. Phon scale is still a reference to

8 American National Standards Institute.

the physical level scale and another scale, the sone scale, has been defined to be directly proportional to loudness. One sone is defined by the loudness curve at 40 phons, 2 sones is defined by the curve perceived as twice as loud as one sone and so on. The curves obtained are showed in figure 2.3.

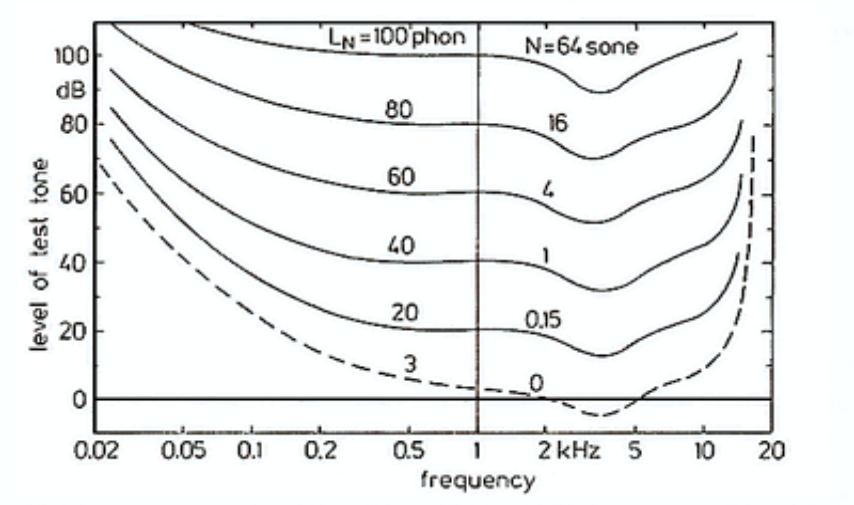



*Figure 2.3: Equal loudness contours. After [Zwicker90].*

The equal loudness contours are standardized as ISO[9] 226 (last revised in 2003).

Pitch is the "attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low" (ANSI, 1973). It is mostly used for harmonic sounds, for which it is related to fundamental frequency. It has been shown that when the fundamental is missing, our auditory system is able to 'reconstruct' it, and the corresponding pitch is perceived (this phenomenon is often referred to as 'the missing fundamental' or 'virtual pitch' ). The equivalent attribute could also apply to noises, which can often be characterized on a scale from low to high frequency. However, 'pitch of noise' has been very little investigated, and the few existing studies were done for very specific sounds [Zwicker90]. Pitch perception is proportional to the logarithm of frequency, i.e. the differences between 200 and 400 Hz tones and between 1000 and 2000 Hz are perceived similarly. A well-known musical pitch scale is the equal temperament scale, in which the pitch is proportional to the logarithm of frequency. A non-musical pitch scale, the mel scale [Stevens37], obtained by experiments similar to those performed for the sone scale, is a scale of pitches judged to be at equal distance one from another. It showed, as illustrated in figure 2.4, that the pitch is not exactly proportional to log-frequency.

9 International Organization for Standardization.

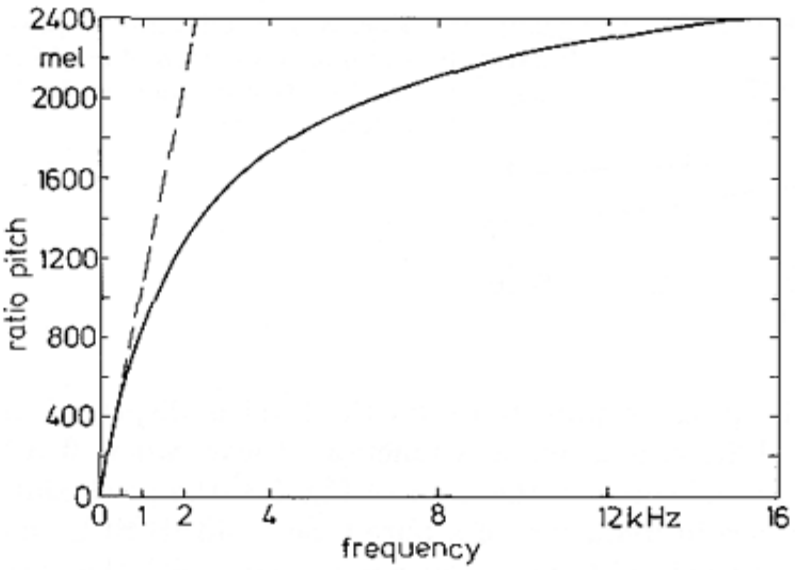



*Figure 2.4: Ratio pitch (mel scale) versus frequency. The dashed line is a linear approximation of the mel scale at low frequency. After [Zwicker90].*

**Timbre**

In the traditional musical notation, the remaining attributes of a sound are indirectly specified by a reference to the instrument and are often referred to as the timbre. Timbre is still an ill-defined and controversial attribute. It is officially defined as "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". (ANSI,1976). This negative definition does not provide any insight into what timbre is. Moreover, it seems to apply only to sounds with pitch, leaving aside most percussive instruments or environmental sounds. As Bregman pointed out, timbre ANSI definition should rather be this: "We do not know how to define timbre, but it is not loudness and it is not pitch" ([Bregman90], p93).

From the numerous definitions found in the literature[10], we can see that timbre is generally associated to two different concepts, related to two listening modes described in section 2.2.1. In the context of a natural, *everyday* listening, timbre is defined as a property of the source, i.e. a set of characteristics that allows identifying

10 A review of definitions of timbre collected in the literature can be found at http://www.zainea.com/timbre.htm.

it[11]. However, analysing sounds from a given musical instrument shows that sound characteristics can be different for different notes (see Schaeffer's experiments on piano in [Schaef66], p234). As Schaeffer noticed, one could identify a specific timbre for each sound that can be produced by a given object, and some patterns of timbre transformation across the register of an instrument could even contribute to its identification[12]. In the context of a *musical* listening, timbre is defined as a property of the sound which allows characterizing it, without any reference to the source. According to that definition, the sounds from two different piano notes, for instance, have different, though quite similar, timbres.

A lot of research has been done in order to understand the underlying qualities of timbre, typically aiming at identifying the physical attributes allowing automatic musical instrument classification (see a review in [Herrera03]). More recent works have investigated timbre for environmental sounds, such as dog barks, bouncing balls or creaking doors) (see a review in [Gygi01]). Although these studies mostly deal with source recognition and are limited to specific sounds, they provide some dimensions of the general perceptual space we are looking for.

Musical sounds

Most experiments on timbre focused on harmonic musical instruments. Since Helmholtz pioneer study on sound analysis ([Helm54]) and until the middle of the 20th century, timbre was very much related to the relative amplitude of the partials of harmonic sounds. In the 60's, some experiments showed that the temporal envelope, especially the attack, also contributed to the perceptual identity of a sound [Schaef66] [Berger64]. Smoothing, for instance, the attack of a sound from percussive instruments or high piano notes modify noticeably its timbre [Schaef66]. Since the 70's, most timbre studies have been done using subjective similarity judgments in pairs of sounds (equalized in pitch, loudness and duration), analysed by multi-dimensional scaling (MDS) techniques. MDS allow « determining the Euclidean space (in an appropriate number of dimensions) within which different timbres can be ordered such that the distances separating them correspond as much as possible to listeners' judgments of their relative dissimilarities. This representation is called "timbre space" and the axes are interpreted as being the perceptual dimensions of timbre » [Donnadieu94]. One of the first experiment using such methodology was led by Grey, who found a 3D timbre space for sounds from 16 wind and string instruments [Grey77]. The first dimension was related to the spectral centroid (the center of gravity of the spectrum, often referred to as the brightness), the second dimension to the amount of synchronicity of partials behaviour (i.e. the amount of

11 This definition is similar to that of ecologists' structural invariants (see footnote 5)

12 Recent experiments on how sound timbre is perceived across the whole register of an instrument and the possible role of timbre transformation patterns in instrument identification are described in [Handel04].

spectral fluctuation) and the third one was found to be correlated to the steepness of the attack. Since then many experiments of the same type derived similar results [Wessel79] [Krum89] [Krim94] [McAd95]. In similar experiments on timbre of percussive sounds, Lakatos found a 2D perceptual space, with the first dimension best explained by the steepness of the attack and the temporal energy distribution and the second dimension related to the brightness [Laka00]. A typical timbre space is shown in figure 2.6.

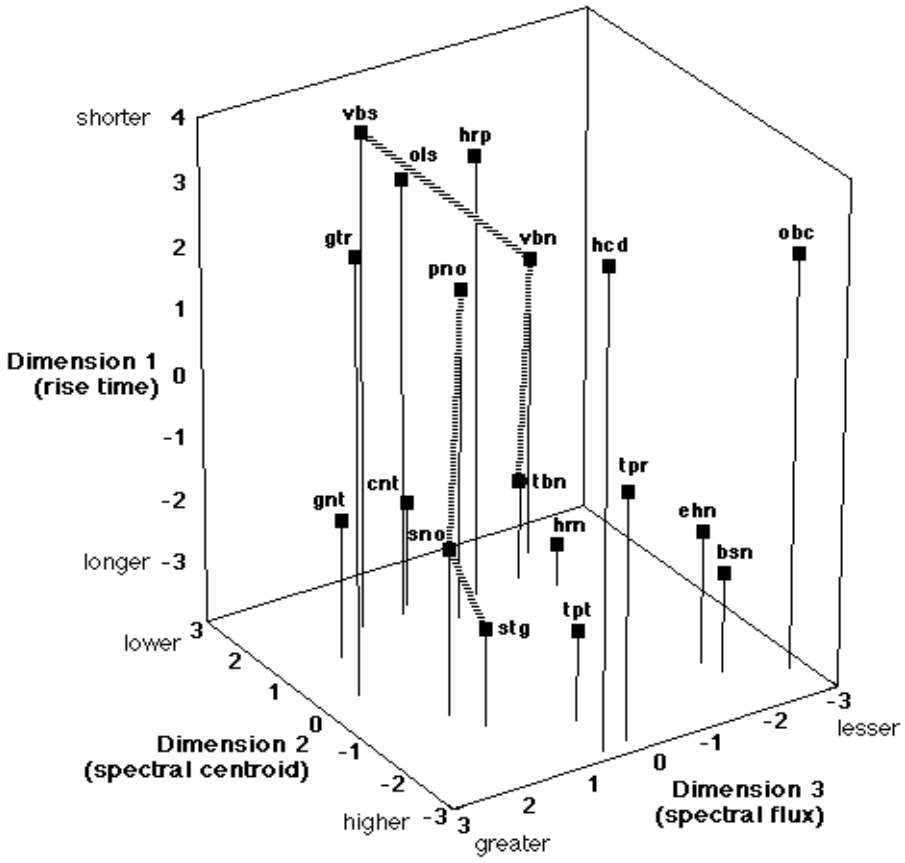



*Figure 2.6: Timbre space derived from similarity judgments of pairs of sounds and MDS analysis in [McAd95].*

Studies on musical sounds seem to agree on quite simple timbre space. However, these kind of sounds represent a very limited sub-space in of a general perceptual space, and studies on non-musical sounds will help complete it.

Non musical sounds

Von Bismarck used subjective judgments on 30 bipolar scales (e.g. low-high, dull-sharp...) to study salient dimensions of laboratory-generated sounds including noises and harmonic sounds with various spectral envelopes [Bism74]. In this method, the scales that do not provide any relevant information (scales equally rated for all sounds or inconsistently rated) are eliminated and scales that are strongly correlated are grouped. The first salient dimension found by von Bismarck was mainly related to the verbal scale referring to the 'sharpness' of a sound and was correlated to the spectral centroid (same as brightness). A second important dimension was found to be associated to the sound 'compactness' and correlated to some measure of noisiness. From experiments on natural sounds from sonar recordings, Howards [Howard77] derived a two-dimensional MDS solution corresponding to the spectral envelope shape and to the amount of low-frequency (<1Hz) periodicity. In her study on the perception of environmental sounds, Van Derveer (1979) found that the grouping by perceptual similarity of sounds from events such as hammering, knocking or paper crumpling were mostly based on the similarity of the temporal pattern of sounds (continuous, repetitive, percussive...) [Vander79]. Another experiment, performed by Bjork and based on the subjective ratings of environmental sounds according to 24 bipolar scales, yielded two salient dimensions, one related to a tense-relaxed scale, which correlated with a measure of roughness, and the other related to a sharpness scale, which correlated with a measure of frequency content, on a scale from low to high [Bjork85]. Recent experiments based on MDS analysis of 100 environmental sounds confirmed the salience of noisiness and temporal pattern (rhythmicity and periodicity) in similarity judgments [Gygi00].

**Roughness**

Roughness, by analogy to the sense of touch, is the attribute related to the perception of short irregularities in a sound. Though it is rarely used or detected in timbre studies[13], it clearly can allow discriminating two sounds, and should then be taken into account in a global perceptual space. Roughness was first identified and defined by Helmholtz after some experiments on the perception of two simultaneous pure sines. By changing the frequency difference ($\Delta f$) between the two tones, he found out that one could identify three 'perceptual zones': Below about $\Delta f$ =10 Hz a listener perceives a tone at the mean frequency modulated in amplitude by a sine at frequency $\Delta f$ (this phenomenon is known as beating), above 10 Hz, modulations cannot be

13 It is, however, often used to evaluate the subjective sound quality. Several publications on this subject can be found in the issue of Acustica/acta acustica, vol. 83(5), 1997.

counted any longer and a new sensation, roughness starts to increase and reaches a maximum at $\Delta f = 70$ Hz [Helm54]. At higher frequency, roughness decreases and the two tones are separately perceived. Further experiments on modulated sines showed that roughness mainly depends on three factors: the modulation depth, the modulation frequency and the carrier frequency [Zwicker90]. Zwicher and Fastl defined a unit, the asper, that corresponds to the roughness of a 60 dB, 1 kHz tone, 100% modulated by a 70 Hz sine [Zwicker90]. The dependence of roughness on carrier frequency and modulation frequency is shown in figure 2.7.

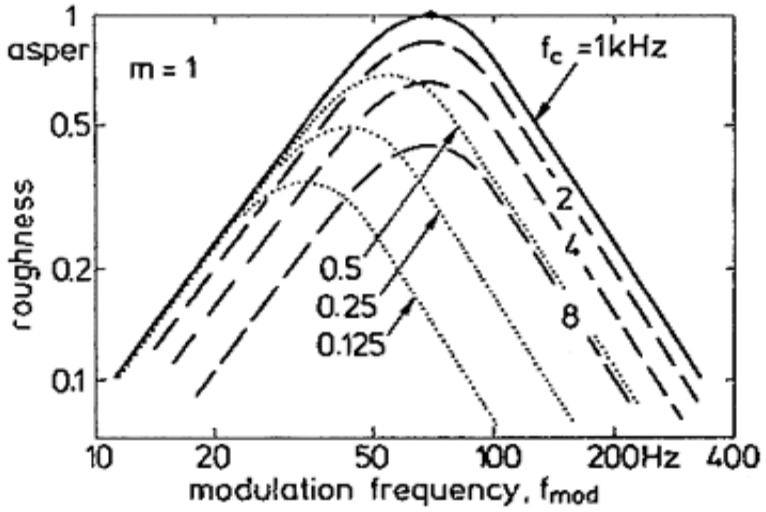



*Figure 2.7: Roughness versus modulation frequency for various pure tones. After [Zwicker90].*

Further research on more complex sounds also showed that global roughness could be predicted from some roughness measure in frequency bands and some measure of temporal envelope coherence across the bands (basically the perceived roughness is higher when the envelopes coherence is high). A review on roughness theories and models, as well as the description of of a model (described in section 3.2) derived from recent experiments, are given in [Press98].

### 2.4.3 Schaeffer's typo-morphology

In his *Traité des objets musicaux* (Treatise on musical object) [Schaef66] (synthesised

and commented by Chion in [Chion83]), Pierre Schaeffer proposed a generalization of what is usually heard as musical sounds (typically notes generated by traditional musical instruments) by considering all kind of *sound objects,* ignoring their origin (electronic sounds, noise, natural sounds, loops…), for which the traditional musical sound representation is way too limited. After some listening experiments, he proposed a general sound classification (*typology*) according to some *morphological* criteria, in order to build a *solfège of musical objects*, foundation of *musique concrète*[14].

**Sound object**

In his research for a generalised solfège, Schaeffer defined the *sound object* as the element of study. Sound objects are the correlates of a *reduced listening* (or, according, to Gaver, a *musical listening* [Gaver93]), during which sounds are listened to for their intrinsic perceptual qualities, independently from their meaning or their origin. In a sound stream, any entity perceived as having its own internal properties and rules is considered as a sound object. In a piece of music, for example, a sequence of notes can be perceived as a single entity, i.e. a musical phrase, as well as the succession of smaller sound objects, i.e. the notes themselves. The sound object is the result of a particular intention, for which any sound is listened to the same way, providing a good basis for general perceptual sound description.

**Typo-morphology**

In order to describe and classify sound objects, Schaeffer defined a *typo-morphology*, in which sound objects are categorized into a typology based on some perceptual attributes, called *morphological* criteria. The building of this typo-morphology is based on the pair of criteria *form/matter*. The sound matter is defined by Schaeffer as *what we would hear if we could freeze the sound* (then mainly –but not only- related to spectral distribution), while the form is related to the time evolution of this matter.

These criteria were studied by listening to sounds with fixed matter, in order to focus on the form, and sounds with fixed form to study the matter. Varying sounds, in which both the form and the matter vary, are also studied through the *variation* criteria. By refining the rough sound description and classification he obtained through these three criteria, Schaeffer defined seven morphological criteria related to different perceptual dimensions emerging from reduced listening:

14 'Musique concrète' (concrete music) composition starts from concrete sound material arranged in such ways that some music emerges from it, as opposed to abstract music, which starts from an abstract representation, the score, and is played later.

Matter criteria

- *Masse (Mass)*: related to the perception of the 'pitchness' of a sound (i.e. a scale from noisy to pitched), and then to its spectral distribution. Schaeffer defines four types of mass: *pitched* (fixed mass and identifiable pitch), *complex* (fixed mass and non-identifiable pitch), *varying* (*pitched-varying* or *complex-varying*, for small or organized variation) and *nondescript* (excessive and unpredictable variation).
- *Timbre harmonic (Harmonic timbre)*: *the more or less diffuse halo associated to the mass and more generally what allows describing it* ([Schaef66], p516). It relates to a finer characterisation of the mass, often described by analogy to vision: bright/mat, round/sharp…
- *Grain (Grain)*: roughness, defined as the micro structure of sound matter, such as the rubbing of a bow. Even though it has a temporal dimension, it is a matter criterion. It is divided into three types: *resonance* grain, for non-sustained sounds (e.g. cymbal resonance), *rubbing* grain, for sustained sounds (e.g. bow or breath sounds) and *iteration* grain, for iterative sounds (e.g. drum roll).

Form criteria

- *Dynamique (Dynamics)*: criterion related to the shape of the amplitude envelope. Schaeffer distinguished several types (e.g. unvarying, impulsive...), as well as several types of attack (smooth, steep...).
- *Allure (Pace)*: amplitude or frequency modulation. Three types: mechanical (very regular), lively ("flexible periodicity, revealing a living being") and natural (unpredictable).

Variation criteria

- *Profile mélodique (Melodic profile)*: related to the variation of the pitch. Schaeffer defined nine types, according to three variation types ('imperfect stability', continuous -e.g. a glissando- and discontinuous -e.g. a piano phrase- variation) and to three variation speeds (slow, medium and fast).
- *Profile de masse (Mass profile)*: variation within the mass. Schaeffer defined several typical mass variations, e.g. 'pitch to complex' or 'thin to thick'.

The complete description scheme was summarized by Schaeffer in a table shown in Appendix 1. A similar classification system, though simplified, is used by Schafer to describe the sound events occurring in natural sound scenes, called *soundscapes* by the author [Schafer77].

### 2.4.4 Perceptual categories

The attributes reviewed in chapter 2 can be seen as dimensions of a generic perceptual space in which any sound can be represented by a set of coordinates specified by a numerical value (e.g. roughness = 0.5 asper, or pitch = 400 Hz) or a type (e.g. pitch variation = 'continuous', or amplitude envelope = 'impulsive'). In specific applications, it can be useful to categorize sounds by identifying sub-spaces corresponding to some types of sounds sharing a set of properties. Well known perceptual categories are onomatopoeia, which consist in using words that imitate the sounds they denote (e.g. bang, buzz, beep, meow...), but all sounds cannot be described that way. The only attempt to provide a classification system applicable to any sound is Schaeffer's typology, based on the morphological characteristics (see 2.3.2) and the length of the sound. For instance, sounds with unvarying matter and form (e.g. white noise or sustained organ note) are called 'homogeneous' sounds, and a typical musical sequence of note is called a 'group' [Schaef66] [Chion83].

### 2.4.5 Summary

The perceptual criteria described in this chapter as well as some proposals to characterize them are summarized in table 2.1.

| *Criteria* | *Brief description* | *Characterization (proposals)* |
|---|---|---|
| **Subjective duration** | Perceived duration of a sounds. | Categories: e.g. short, medium, long. |
| **Loudness** | "(...) attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from soft to loud." (ANSI 1973) | Categories: e.g. soft, medium and loud. |
| **Pitch** | "(...) attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low." (ANSI, 1973) | Continuous frequency scale (in Mel or Hz), or discrete values (e.g. musical note) when possible for harmonic sounds.<br>Categories: e.g. low, mid or high frequency for noisy sounds.<br>+temporal pattern characterization, e.g. , ascending descending... |
| **Pitchness** | Scale from noise to pitched. This attribute is related to the pattern of the frequency partials distribution. | Categories: e.g. noise, pitched ..<br>+temporal pattern characterization, e.g. pitched to noisy... |
| **Roughness** | Attribute related to the perception of short amplitude envelope irregularities in a sound. | Categories: e.g. rough, medium, smooth.<br>+temporal pattern characterization, e.g. increasing, decreasing... |
| **Dynamic profile** | Shape of the amplitude envelope. | Categories: e.g. unvarying, impulsive, crescendo... |
| **Attack** | Attribute related to the perception of the temporal pattern of the attack | Categories: e.g. smooth, steep, straight... |
| **Brightness** | Attribute related to the perception of the spectral centre of gravity. | Categories: e.g. low, medium, high.<br>+temporal pattern characterization, e.g. increasing, decreasing... |
| **Spectral fluctuation** | Attribute related to the perception of short spectral fluctuation | |
| **Others ('Metallicness', Richness...)** | To be defined... | |

*Table 2.1Summary table of perceptual description criteria*

## 2.5 Discussion

Our objective is to define a perceptual description scheme as complete as possible. Some experiments on timbre have led to the identification of some perceptual dimensions for specific sounds (musical sounds and environmental sounds), and a more general timbre space can be derived by combining them. The early work on sound objects done by Pierre Schaeffer was the first attempt to provide some general description criteria and, though his typo-morphology was defined rather intuitively, by transforming and listening to all kinds of sounds, it includes most of the dimensions derived from timbre studies on both musical and environmental sounds. The morphological criteria based on the pitch and pitchness trajectories (melodic and pitchness profile) have not been identified by these studies as salient perceptual dimensions, but it seems clear that they all help characterizing sounds and that should be part of a general perceptual description scheme. Table 2.1 provides a basis for such a description scheme, in which any sounds would be described for each criteria by a numerical value or a typical category.

# 3 Computational perceptual sound description

In this chapter we review the models aiming at mimicking some features of auditory perception.

## 3.1 Modelling auditory perception

Modelling auditory perception is a difficult and multi-disciplinary task. Perceptual criteria cannot be easily studied directly from the ear physiology. Dimensions of our auditory perceptual space should first be identified (see Chapter 2) and specifically investigated by combining physiological and cognitive data as well as psychoacoustical data derived from listening experiments. While low-level processes, such as the frequency decomposition performed by the cochlea, are quite well understood and have been extensively described (e.g. see [Zwicker90] or [Green01]), current knowledge in auditory perception is limited and does not allow to construct accurate models of our perceptual experience of sound. Numerous models have, however, been proposed by the research community, for various purposes, ranging from complex systems modelling as far as possible the human auditory system, to more application-oriented models, aiming at providing simple and efficient methods to extract perceptual features for specific sounds.

## 3.2 Models of perceptual attributes of sound

**Basic modelling of the peripheral auditory system**

Since all high-level auditory processes giving rise to mental representations are preceded by a pre-processing of the acoustical signal in the ear, designing a computational model of some perceptual criteria often require to mimic all or some part of this pre-processing. Many studies on peripheral auditory system modelling, derived from psychoacoustical data or even from measurements performed on animals, can be found in the literature (see e.g. [Green01] [Zwicker90] or [Kabal02]). A schematic view of the ear is shown in figure 3.1.

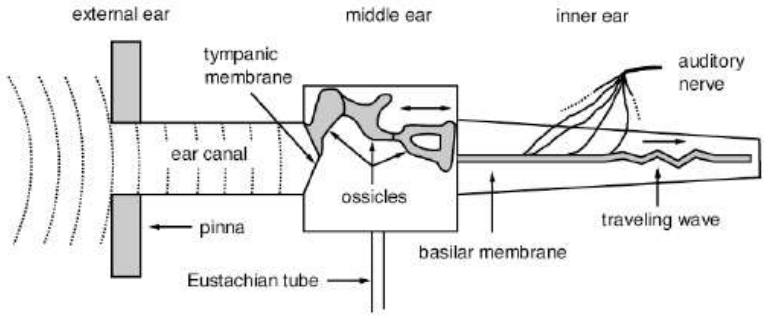



*Figure 3.1: Schematic view of the peripheral auditory system. After M. Karjalainen's lecture material. Reprinted with permission of the author.*

The outer and middle ears can be modeled by a simple filter, which transfer function $A(f_{kHz})$ is given by equation 3.1 and shown in figure 3.2 [Kabal02].

$$A_{dB}\left(f_{kHz}\right) = -20184\left(f/1000\right)^{0.8} + 6.5\,\mathrm{e}^{-0.6\left(f/1000-3.3\right)^2} - 0.001\left(f/1000\right)^{3.6} \qquad (3.1)$$

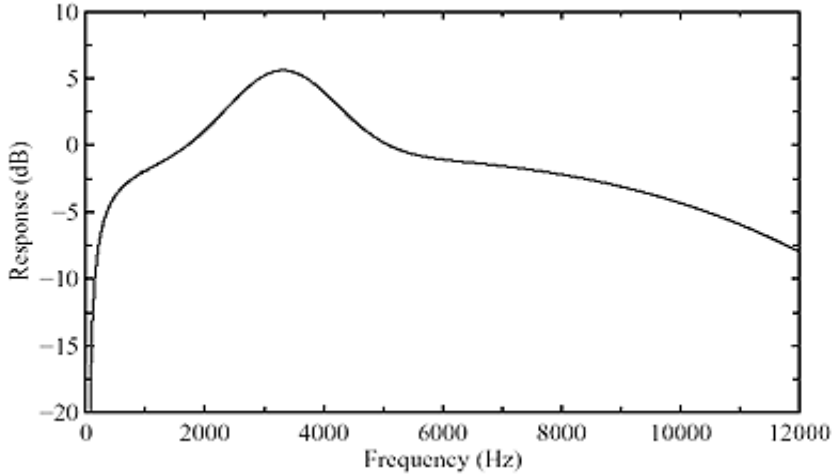



*Figure 3.2: Transfer function of the outer and middle ear. After [Kabal02].*

The most important feature of the peripheral ear processing is the decomposition of the signal into frequency bands, referred to as critical bands, performed by the cochlea. Some experiments showed that this frequency decomposition could be approximated by the *bark scale,* defined between 0 and 15.5 kHz by 24 critical bands of center frequency $f_c$, with bandwidth equal to 100 Hz below 500 Hz and to 0.2 $f_c$

above [Zwicker90]. Another measure of the width of critical bands is the Equivalent Rectangular Bandwidth (ERB) [Moore83], given in [Moore96] as a function of center frequency $f_c$ by equation 3.2.

$$ERB = 24.7 + 0.108 \times f_c \tag{3.2}$$

**Loudness**

The equal-loudness contours described in 2.3.2, based on the perception of pure tones, are not sufficient for estimating the loudness of complex sounds. Further experiments showed that the loudness produced by two tones of equal level corresponds to the addition of the loudnesses of each tone when their frequency difference is large, but that they influence each other and produce a smaller loudness when this difference is smaller than a critical bandwidth. Loudness models, based on recent psychoacoustic data, include outer and middle ear filtering models and estimate the loudnesses in each critical band [Zwicker90] [Moore97][15], as illustrated in figure 3.3.

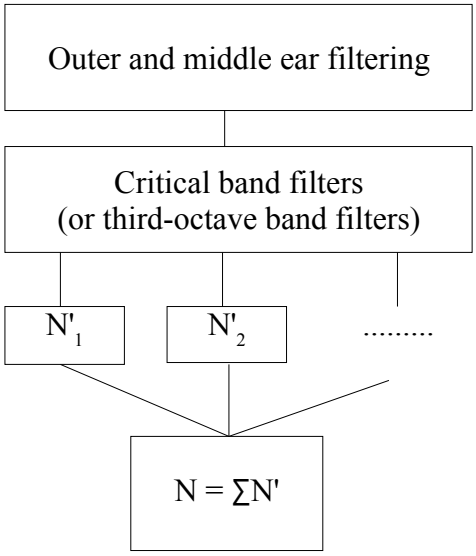



*Figure 3.3: Block diagram of loudness models proposed in [Zwicker90] [Moore97].*

15 Computer models based on Moore et Al.'s loudness model are available at http://hearing.psychol.cam.ac.uk/Demos/demos.html or in Psysound computer program for psychoacoustical analysis [Cabrera99].

Each specific loudness N' is estimated assuming a relationships $N' \sim E^{\alpha}$, where E stands for the excitation level in the band and $\alpha$ a coefficient reflecting a compression performed by the auditory system, and an additional excitation due to internal noise in the ear. A review of loudness models is given in [Appell02].

**Pitch**

Models of pitch perception basically consist in estimating the fundamental frequency ($F_0$), its physical counterpart, of harmonic sounds. Different pitch estimation methods have been investigated, which have been recently reviewed by Klapuri in [Klap04]. He proposed to classify existing algorithms in two groups, according to whether they are based on the location of the partials (*spectral place*) or on the intervals between pairs of partials (*spectral interval*).

Spectral place based methods

A widely used method for F0 estimation, in the time-domain, is to look for the predominant periodicity in the waveform (e.g. [Brown91] or [Talkin95]). A simple measure of periodicities is the autocorrelation function (ACF) r(n), given for a discrete signal x(n) of length K by

$$r\left(n\right)=\frac{1}{K}\sum_{k=0}^{K-n-1}x\left(k\right)\times x\left(k+n\right) \tag{3.3}$$

The highest peak in the ACF is taken as the period of the fundamental frequency. The ACF can also be computed in the frequency domain as

$$r\left(n\right)=IDFT\left(\left|DFT\left[x\left(n\right)\right]\right|^2\right) \tag{3.4}$$

where IDFT stands for the Inverse Discrete Fourier Transform and DFT for the Discrete Fourier Transform.

For real signal, this can be rewritten as a function of the magnitude of the Fourier Transform X(k) of the signal as

$$r\left(n\right)=\frac{1}{K}\sum_{k=0}^{K-1}\left[\cos\left(\frac{2\,\pi\,n\,k}{K}\right)\left|X\left(k\right)\right|^{2}\right] \tag{3.5}$$

which shows that the ACF actually corresponds to a weighting that emphasizes partials at harmonic location [Klap04]. A similar method, based on the cepstrum, is obtained by replacing the squared spectral magnitude in equation 3.4 by the log spectral spectral magnitude [Noll67]. Another method, referred to by Klapuri as *harmonic pattern matching*, is to compare directly the frequency spectrum to ideal harmonic frequency patterns and to select the best match (e.g. [Brown92], [Dov93] or [Maher94].

The main drawback of methods based on the location of the partials $F_n$ is that they do not handle well inharmonicity. High order partials of piano sounds, for instance, are shifted towards in frequency so that the deviation from a perfect harmonic pattern ($F_n = nF_0$) increases with the order of the partial. In that case, methods based on partials intervals, more stable than partials locations, perform better.

Spectral interval based methods

Since perfect harmonic spectra have periodic partials distributions, $F_0$ can also be estimated from autocorrelation of the magnitude spectrum (i.e. the time-domain signal is replaced by the frequency spectrum in equation 3.3) (e.g. [Lahat87] or [Kun96]). As Klapuri pointed out, the information provided by this method is quite different from time-domain ACF since any interval close to the fundamental frequency increases the autocorrelation at the corresponding lag.

Envelope periodicity

A very different approach, based on the beating phenomenon described in 2.4.2 about experiments on roughness, was proposed in [Meddis97]. Since each pair of partials ($F_n$, $F_m$) produces an amplitude modulation of frequency $\Delta F = |F_n - F_m|$ in the temporal envelope, a predominant envelope periodicity should be observed in harmonic signals at $F_0$. Meddis et Al. used this principle in an attempt to model human pitch perception: The signal is first filtered by a filterbank modeling the cochlea. The temporal envelope of the output of each filter is then extracted and analysed by autocorrelation. Autocorrelations are summed across frequency channels in a *summary autocorrelation function* and the highest peak is taken as the pitch period. This method demonstrated good performance and was able to reproduce many

features of human pitch perception.

**Pitchness**

All the algorithms described above aims at estimating the pitch of harmonic sounds, typically on a frame-by-frame basis. In general sound analysis one first needs to know whether a given frame is pitched or not in order to avoid inconsistent pitch estimation. In harmonic pattern matching methods, the *pitchness* can be measured simply from the mismatch between the magnitude spectrum of the signal analysed and the ideal harmonic pattern that best matches it. If the mismatch between the analysed signal and the best match is large, the sound is considered as inharmonic. An example of this approach is given in [Cano98]. In [Slaney98], a measure of pitchness, called *pitch salience*, is estimated by the ratio between the amplitude of the highest peak of the ACF and the total power of the signal, given by the amplitude of the ACF at 0. This ratio is close to 1 for harmonic sounds and close to 0 for non-harmonic sounds, which are not periodic and exhibit then no peak in the ACF. A measure of noisiness, the opposite scale, can also be estimated by taking the ratio of the noisy component energy to total energy in a sinusoidal plus noise analysis [Serra98].

**Timbre**

In 2.3.2 we reviewed some studies that allowed identifying some dimensions of a perceptual timbre space. Recent studies investigated the physical correlates of these dimensions for sounds from musical instruments [Krim94] [Peeters00] or from environmental sounds [Gygi00] [Bone01]. The physical features found for each dimension can be classified as temporal, spectral or spectro-temporal features.

Temporal features

A dimension of all the timbre spaces reviewed in chapter 2 was shown to be strongly correlated to temporal characteristics: steepness of the attack for sounds from harmonic sounds, steepness of the attack and temporal energy distribution for percussive musical sounds and global temporal pattern (continuous, repetitive, percussive...) as well as some measure of periodicity and rhythmicity for environmental sounds. The best physical measures of the steepness of the attack and the temporal energy distribution were found to be, respectively, the logarithm of the attack time and the temporal centroid. The attack time is computed from the time at which the signal reaches a given threshold to the time at which the signal reaches its maximum or its sustained part. The temporal centroid is the center of gravity of the temporal envelope A(n) of length N, given by equation 3.6.

$$TC = \frac{\sum_{n=0}^{N-1} A(n) \cdot n}{\sum_{n=0}^{N-1} n} \qquad (3.6)$$

The dimension related to rhythmicity and periodicity was found to be correlated to some measure of the amount of silence in the sound and some autocorrelation statistics (number of peaks and maximum value of the peaks) [Gygi00].

Spectral features

Another dimension which seems common to all types of sound is the brightness (or sharpness), correlated to the spectral centroid (computed by replacing the time domain signal by the frequency spectrum magnitude in equation 3.6). A more complex measure of brightness, based on psychoacoustical data, is described in [Zwicker90].

Spectro-temporal features

The third dimension of timbre spaces derived for harmonic sounds is related to spectral fluctuations. Whereas most studies agreed on the physical correlates of the first two dimensions, the last one seems more dependent on the stimulus used. In their study of two timbre space ([Krum89] and [McAd95]), [Krim94] found the third dimension to best correlate to *spectral irregularity* (log of the spectral deviation of component amplitudes from a global spectral envelope derived from a running mean of the amplitudes of three adjacent harmonics) in one case and to *spectral flux* (average of the correlations between amplitude spectra in adjacent time windows) in the other case. A more recent study found the third dimension of the space obtained in [McAd95] to be best described by three features: the *harmonic spectral spread* (the extent of the spectrum energy around the spectral centroid), the *harmonic spectral variation* (the amount of variation of the spectral energy distribution along time) and the *harmonic spectral deviation* (the deviation of the harmonics from a global spectral envelope) [Peeters00]. These features were integrated in the MPEG 7 audio standard described in 3.3.

Some measures of pitchness and roughness, also found to be salient dimensions of timbre spaces derived from environmental sounds, are described in specific paragraphs.

**Roughness**

We saw in section 2.4.2 that roughness was due to amplitude modulations in a given frequency range. Several models have been developed according to two different approaches. Frequency-domain models indirectly measure amplitude modulations by estimating the beating phenomenon between pairs of partials, while time-domain models directly estimate the amount of amplitude modulation in the roughness range. Many experiments on this phenomenon were done using simple stimuli (typically pairs of tones) in a frequency range smaller than a critical bandwidth and then only led to models of partial roughness (roughness in a critical band). In a review of experiments and models of roughness of complex sounds, Pressnitzer showed that global roughness could not be estimated by simply adding up partial roughnesses and that more complex mechanisms should be taken into account [Press98]. As he pointed out, "this result is intuitively known: non-modulated white noise produces almost no roughness, although summing up the fluctuations it causes in each critical band should results in high roughness." Early experiments, completed by Pressnitzer, showed that the main factor in partial roughnesses addition was the coherence of the envelopes of all frequency bands: Sounds exhibiting modulations in the roughness range in several critical bands produce high roughness when these modulations are in phase whereas they produce low roughness when they are not (this is the case in white noise). From these observations, Pressnitzer designed a model based on the temporal approach, described in figure 3.4.

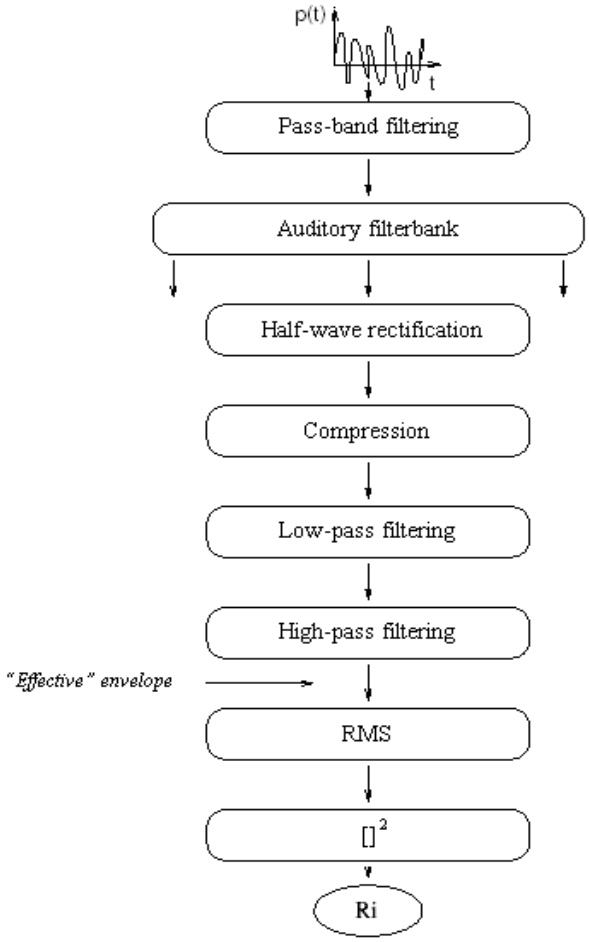



*Figure 3.4: Block diagram of Pressnitzer's partial roughness model [Press98].*

The signal is first pass-band filtered between 450 and 1000 Hz and decomposed into critical bands using a filterbank modeling the cochlea. The amplitude envelope of each band is then extracted by half-wave rectification ($x(n)=0$ if $x(n)<0$) and low-pass filtering and compressed by taking the square root. The amount of modulation in the roughness range is extracted by a low pass filter of order 2 (12dB/octave) with cut-off frequency at 70 Hz followed by a high pass filter of the same order and at the same cut-off frequency. Partial roughnesses $r_i$ are then computed by computing the squared RMS value of the resulting signal, called *'effective' envelope*, in each band. Global roughness R is obtained by combining the results for N bands, as shown in equation 3.7.

$$R = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ j \geq i}}^{N} c_{i,j}\left(r_i + r_j\right) \qquad (3.7)$$

$c_{i,j}$ stands for the correlation between the effective envelopes of channels i and j and $r_i$ for the partial roughness of channel i.

## 3.3 MPEG7 Audio

The MPEG-7 is the first attempt to provide a standard to audio content description. It has been developed by the MPEG (Motion Picture Expert Group) since 1996 to provide a set of description tools aiming at managing and retrieving multimedia (audio and visual) content on the web or in databases [Mar03]. The descriptors defined in the MPEG 7 audio framework are divided into several types:

- Silence.
- Basic: simple descriptors from the waveform (e.g. minimum, maximum) and instantaneous power.
- Basic spectral: simple descriptors derived from the frequency spectrum: spectral flatness, centroid and spread.
- Signal parameters: fundamental frequency and harmonicity (a measure of pitchness, see 3.2).
- Spectral basis: descriptors representing low-dimensional projections of the spectrum (singular value decomposition, principal component analysis....).
- Timbral[16] temporal: log-attack time and temporal centroid.
- Timbral spectral: harmonic spectral centroid, harmonic spectral deviation, harmonic spectral spread, harmonic spectral deviation and harmonic spectral spread

These low-level descriptors (LLDs) are mainly used as features for higher level analyses performed by specific tools (e.g. musical instrument timbre description,

16 Timbral descriptors are the physical correlates of some dimensions of timbre spaces derived from subjective experiments (see 2.4 and 3.2).

melody description, general sound recognition) also included in the MPEG 7 audio framework. Examples of MPEG 7-based applications are given in [Peeters00] and [Gomez03b].

## 3.4 Discussion

We showed in this chapter that most of the perceptual features described in chapter 2 can be analysed by some computational models, derived from physiological or psychoacoustical data, which could be combined to extract automatically a complete perceptual representation. The recent standard for audio description, the MPEG-7 audio, includes some perceptual descriptors specific to harmonic and percussive musical sounds, but lacks features for more general perceptual sound description. For instance, there are no features allowing to characterise the temporal pattern (except the attack) or the roughness of a sound, that was both found to be salient dimensions of timbre spaces derived from listening experiments on environmental sounds. In next chapter we present our first attempt to design a system for automatically describing sound according to a description scheme, such as the MPEG-7 audio, based on general perceptual criteria.

# 4 Computational morphological sound description

Part of this chapter was published in [Ric03a], [Ric03b] and [Ric04].

We describe in this chapter a system that allows to automatically generate a representation based on Schaeffer's typo-morphology. This system extracts specific features that are used for characterizing each dimension by a numerical value or a class. Some are physical correlates of timbre dimensions or derived from perceptual models described in Chapter 3, and new features have been created when necessary.

## 4.1 Simplified morphological description scheme

Schaeffer's typo-morphology, shown in Appendix 1, is quite complex and some criteria or classes are not well-defined (e.g. harmonic timbre) so we have not attempted to exactly reproduce it in our automatic morphological description system. We started from a very simple description scheme (i.e. a set of descriptors) based on criteria that have been found to be salient dimensions of timbre spaces derived from studies on environmental sounds (see 2.4), namely dynamic profile and pitchness, and on pitchness profile (which was easily computable from pitchness). As we progressed, we have completed the scheme by adding more criteria, classes, sub-classes or descriptors specified by numerical values that have also been found to be important for discriminating sounds and/or that were parts of Schaeffer's typo-morphology. The current morphological description scheme is shown in table 4.1.

| ***Morphological Criteria*** | ***Classes, sub-classes and additional descriptors*** | |
|---|---|---|
| Dynamic profile | Unvarying | |
| | *Varying* | *Impulsive*<br>*Iterative* (several transients) + *periodicity value* + *velocity value*<br>*Crescendo*<br>*Decrescendo*<br>*Delta* (i.e. Crescendo-Decrescendo)<br>*Other* |
| Pitchness | *Pitched* (one predominant pitch)<br>*Complex* (simultaneous pitched components or simultaneous or sequential pitched and noisy components)<br>*Unpitched (noise)* | |
| Pitchness profile | *Unvarying*<br>*Varying* (e.g. from noisy to pitched) | |
| Pitch profile<br>(only specified for pitched sounds) | *Unvarying + pitch value* | |
| | *Varying* | Type of variations:<br><br>*Continuous* (e.g. siren)<br>*Stepped* (e.g. piano phrase) |
| | | *Ascending*<br>*Descending*<br>*Delta* (i.e. *Ascending-Descending*)<br>*Inverse delta*<br>*Other* |
| Harmonic timbre | Brightness value | |
| (Roughness) | Roughness value | |

*Table 4.1: Current scheme of the automatic morphological description system.*

This description scheme includes the following criteria:

- Dynamic profile: describes the shape of the temporal envelope. For iterative sounds (sounds having several transients), some numerical values of velocity (transients per second) and periodicity are also specified.
- Pitchness profile: discriminates sounds with one predominant pitch, sounds with several pitches and sounds with no pitch.
- Pitchness profile: describes whether the pitchness is constant or varies in function of time (in that case, pitchness is the mean value).
- Pitch profile: describe the variation of the pitch, only specified for pitched sounds. For sounds with unvarying pitch, the pitch is given. Pitch-varying sounds sounds are classified according to the type of variation (continuous or stepped) as well as the global envelope of the pitch (e.g. ascending, descending...).
- Harmonic timbre criteria, specified by a numerical value of brightness.
- Roughness, described by a numerical value.

According to these criteria, a piano phrase of several low-frequency ascending notes, for instance, would be described as follows: dynamic profile = 'iterative', pitchness profile = 'pitched', pitchness profile = 'unvarying', pitch variation type = 'varying-stepped', pitch envelope = 'ascending', a low brightness value and a low roughness value.

## 4.2 Features extractions

For each morphological criteria, specific low-level descriptors (LLDs) are computed and used for the classification. The block diagram of our system is shown in figure 4.1.

Each sound file is considered as a single sound object, i.e. no segmentation is performed. If the energy envelope has no value greater than a given threshold, the sound is just labelled as silence. If the signal is not silent, it is filtered by the model of the outer and middle ear given in equation 3.1 and three independent modules extract the LLDs.

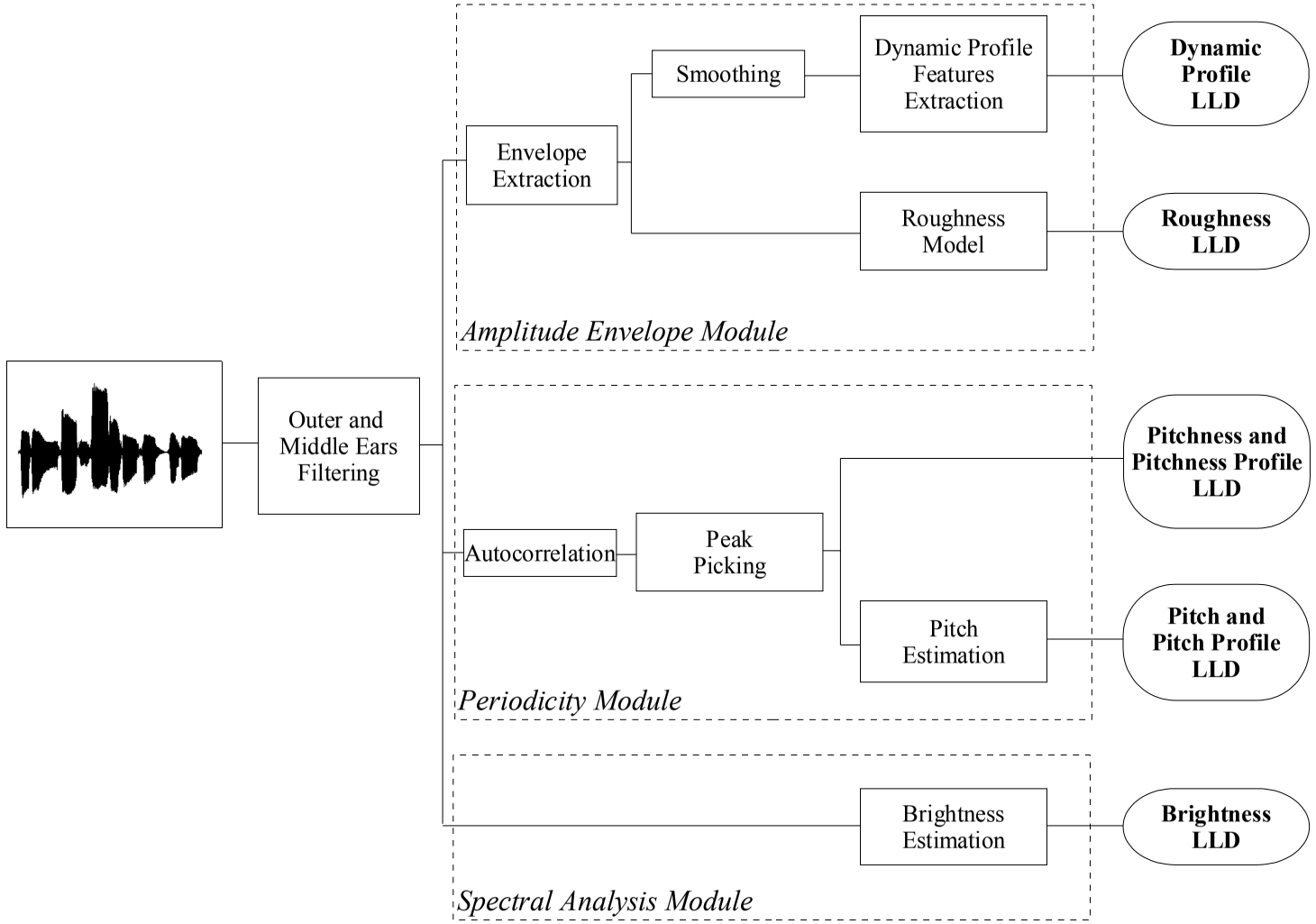



*Figure 4.1: Block diagram of the automatic morphological description system.*

### 4.2.1 Amplitude envelope module

The amplitude envelope is estimated by applying two cascaded one-pole low-pass filters with a 2 ms time constant on the full-wave rectified (absolute value) waveform.

This filter has a -3 dB cut-off frequency of 80 Hz and a roll-off of -12 dB per octave at higher frequencies, so that the range of amplitude modulation of the envelope producing roughness is preserved (see figure 2.7). Short-term roughness is estimated by computing the ratio of the energy between 20 and 200 Hz to the total energy of the magnitude spectrum of the envelope in 250 ms rectangular windows. Total roughness is obtained by averaging this ratio weighted by the envelope energy over the sound. This very simple model was implemented recently for testing purpose. It does not provide accurate estimation of roughness but it does discriminate 'extreme' sounds, i.e. very rough and very smooth sounds.

The LLDs used to classify sounds into one of the dynamic profile classes, shown in table 4.1, are all computed from the log-amplitude envelope (for rough modelling of loudness perception) smoothed by a 60 ms half-hanning window. The other module based on amplitude envelope analysis computes the LLD allowing to classify sounds into one of the Dynamic Profile classes showed in table 6. The LLD were chosen intuitively according to the specificities of each class. For instance we assumed that the centre of gravity of the envelope (the temporal centroid) would allow discriminating between crescendo, unvarying and decrescendo, according to whether it is rather at the left, at the middle or at the right of the envelope, and that a crescendo sound would have a high amplitude derivative average and a maximum value close to the beginning. A simple onset detection, based on the envelope derivative, is used to detect iterative sounds (sounds with several transients). The half-hanning window used to smooth the envelope preserves sudden changes, from which the onsets are detected, but masks rapid modulations, which could produce false onsets detection. For iterative sounds, some estimations of onset velocity and periodicity are also computed. The velocity is a simple measure of the number of onsets per second, and the periodicity is estimated from the histogram of the intervals between each pairs of onsets. Histogram values are first added over a Gaussian window (which size is fixed at the smallest interval value), and the periodicity is estimated by the ratio of the highest histogram peak value to the number of onsets. The periodicity value is close to 1 (exactly (N-1)/N for N onsets) for perfectly periodic onsets, and smaller for non-periodic onsets.

The initial intuitive selection of descriptors was modified according to the classification performance (see 6.3).

### 4.2.2 Periodicity module

Pitchness and pitch estimation is based on the autocorrelation method (see 3.2). Short-term autocorrelations are computed in the frequency domain, as shown in equation 3.4, in 46 ms overlapping hanning windows. Autocorrelations are then unbiased by dividing them by the autocorrelation of the hanning window. In periodic signals, autocorrelation peaks occur at integer multiples of the periods. The amplitude of the peaks is a measure of the pitchness of a sound. Short-term pitchness is computed from pitch salience, defined in [Slaney98] as the ratio of the magnitude of the highest autocorrelation peak to the magnitude of the ACF at 0. We used the average of short-term pitch salience weighted by the energy of the frame to determine the global pitchness of the signal. Highly pitched sounds will have a pitch salience close to 1 (pure tone), unpitched sounds will have a pitch salience close to 0 (white noise). Using this measure to estimate whether a sound is made of several simultaneous pitched components (*complex* sounds) is not mathematically justified since such a sound will exhibit a pitch salience of 1 at the smallest common periodicity of its partials, i.e. at the lag corresponding to the greatest common divisor of the partials frequency (e.g. a sound made of two harmonic series with fundamental frequencies of 200 and 350 Hz will have a pitch salience of 1 at the lag corresponding to 50Hz). However, when the number of simultaneous pitched components is high, the smallest common periodicity tend to be high and is likely to be greater than the higher boundary of our pitch detection range (given below).

The pitchness profile (varying vs. unvarying pitchness) is estimated from pitch salience variance.

The pitch is estimated from the highest peak in the autocorrelation for pitched sounds only (sound with a single pitch). The lower boundary of the pitch detection range is limited by the size of the analysis window (46 ms) to around 50 Hz. The higher boundary is fixed at 5000 Hz, i.e. the peak picking starts at the corresponding lag in the autocorrelation function ($9^{th}$ lag at a sampling rate of 44100 Hz). In order to achieve better pitch estimation, autocorrelation peaks are tracked over time and some post processing correct local errors. The log pitch variance as well as some pitch ‘step’ (discontinuity followed by flat pitch) detection are then used to discriminate between sounds with pitch that does not vary, that varies continuously or that varies by step. The 'shape' of the pitch trajectory (e.g. ascending or descending) is classified using LLD similar to those used for the dynamic profile (e.g. pitch derivative average, position of the maximum pitch...).

Short-term brightness is measured by the spectral centroid of the analysis windows. Global brightness is computed from the average of the short-term spectral centroid weighted by the energy of the frame.

All these functions are implemented in CLAM[17], the C++ Library for Audio and Music developed at the Music Technology Group[18].

The list of the LLD used for morphological classification is given in Appendix 2.

## 4.3 Classification and models evaluation

In order to evaluate our system, we built a small database (around 200 instances for each class) in which sounds were manually labeled according to the classes defined earlier. The database is very varied, containing sounds from musical instruments, environmental sounds, electronic sounds and loops. Classes of each morphological criterion were modeled by C4.5 decision tree[19] [Win92] (the other classifiers that we tried did not perform significantly better and decision trees give a clear view on how features are related to classes) and tested by 10-folds cross-validation[20].

Amplitude envelope were modeled for each dynamic profile class class (one against all), excepted for the *Varying-Other* class, and the models were tested on the whole database, including those labeled as *Varying-Other*. Classification performance was very good (around 98%) for the *Unvarying* and *Varying-Crescendo* classes, as expected because of the specificity of these classes and for *Varying-Impulse* and *Varying-Decrescendo* classes (around 94%, and most misclassified instances in one class was labeled as belonging to the other). The poor performance (88%), compared to the other classes, of *Varying-Delta* amplitude envelope classification might be explained by the high variability of this class (close to crescendo, symmetric or close to decrescendo). The onset detection correctly detected iterative sounds with very good performance[21] (96%).

Pitchness discrimination between *Noisy* and *Pitched* sounds is close to 100%. Some errors occur when the pitch is too low to be detected by our current analysis settings (basically a too short analysis frame size): No peak is detected in the periodogram and the sound is then classified as *Noisy*[22]. The main problem is the classification of

17 http://www.iua.upf.es/mtg/clam/

18 http://www.iua.upf.es/mtg/

19 All the classification tasks were performed using Weka, available at http://www.cs.waikato.ac.nz/~ml/weka/

20 X-fold validation consists in building the models on 90% of the data and testing it on the 10% remaining. The process is performed 10 times on all the possible combinations. The performance given by this method is more realistic than when using exactly the same data set for training and testing.

21 This performance must not be confused with the performance of the onset detection. Here the task was just to estimate whether a sound contains several onsets, even if some are missing, which is much simpler than detecting all the onsets.

22 This will be solved by using an adaptive analysis frame size.

sounds having *Complex* pitchness (simultaneous pitched components, or simultaneous or sequential pitched and noisy components): only 50% are correctly classified, the rest being classified as noisy or pitched, in the same proportion. Two issues should be addressed here: whereas the *Noisy*/*Pitched* sounds discrimination is quite obvious, the perceptual boundary between *Noisy* and *Complex* and that between *Complex* and *Pitched* classes is not clear. Moreover, as we already noticed in 4.2, though the pitch salience is a good feature for detecting complex sounds having simultaneous or sequential pitched and noisy components, it is not very appropriate for detecting those that have several pitched components. In that case, a multi-pitch estimation would be necessary.

Pitchness profile classification performance is 70%, which is not very good for a 2-classes discrimination task. The pitch salience variance isprobably not appropriate or sufficient. It works very well for sounds having a clearly unvarying (e.g. white noise or pure sinusoid) or a clearly varying pitchness (e.g. sequence of white noise and pure sinusoids), but we noticed that many noisy sounds that have a highly varying pitch salience are perceived as unvaryingly noisy. Smoothing the pitch salience did not change the performance.

Pitch profile classification error rates are tightly related to automatic pitch estimation performance. Sounds having an unvarying pitch are correctly classified for 90% of the instances. Varying sounds are further classified as *continuous* or *stepped* with a performance of 85%. When the pitch is correctly estimated, pitch profile classification is almost perfect.

The remaining criteria and classes were implemented recently, and only rough evaluation was performed. In a database of 15 sounds per class, the performance of classification into typical pitch patterns (ascending, dedescending, delta, inverse delta and other) was around 90%. Regarding the roughness, a database containing 20 instances of 'rather rough sounds' and 20 of 'rather smooth sounds' was classified correctly in 85% of the cases. No evaluation was performed for brightness, but the feature we used (the spectral centroid) have been found to be strongly correlated to it in timbre studies, both for musical and environmental sounds (see 2.2.2).

For classes modelled by a combination of features, such as those of the dynamic profile or the pitch profile criteria, the models were obtained by trial and error. We assumed that a poor classification performance was due to a bad choice of features rather than a bad choice of classification algorithm. If two classes were confused, we added features that could help discriminating them better[23].

23 For instance, impulsive sounds and decrescendo sounds were often confused by the classification algorithm. Since the release of impulsive sounds is typically steeper than that of decrescendo sounds, adding some measure of amplitude derivative after the maximum amplitude value increased sensibly the performance.

## 4.4 MPEG7 extension

An extension of the MPEG 7 audio description scheme based on morphological description was built and integrated to an application for content-based retrieval and transformation designed in the context of the CUIDADO project. This tool, called the Sound Palette, is described in [Celma04].

## 4.5 The AudioClas sound search engine

Some morphological descriptors have been integrated as alternative search criteria to a prototype sound search engine designed in the AudioClas[24] project [Cano04]. Some classes of Dynamic Profile, Pitch Profile, Mass (Pitchness) and Mass Profile could be specified for a search only based on perceptual description or for refining a search by traditional keywords related to the origin of the sound. A snapshot of the prototype is shown in figure 4.2.

24 http://www.audioclas.org/

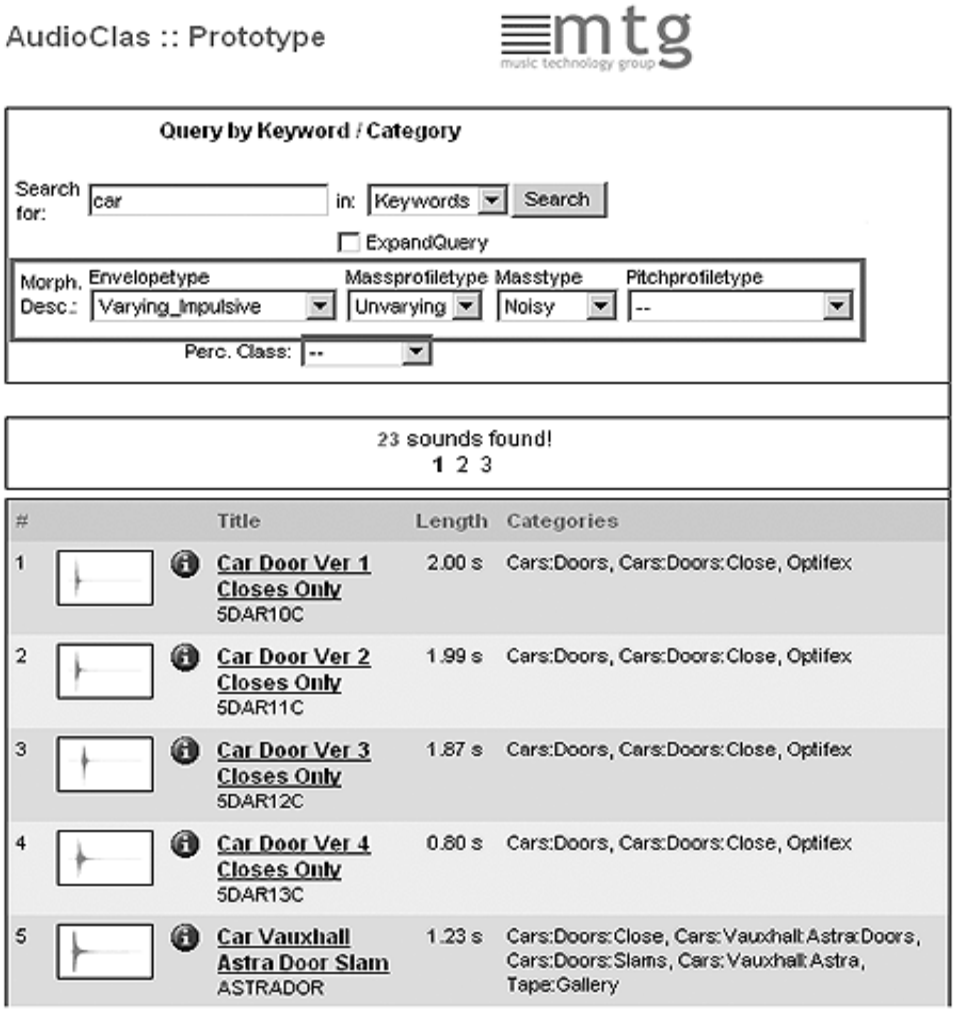



*Figure 4.2: Snapshot of the AudioClas prototype sound search engine. In this example, the original search (using the keyword 'car') retrieved more than 400 sounds. Filtering by morphological criteria refined a lot the search and gave only 23 sounds.*

# 5 Usability evaluation of morphological sound representation for sound retrieval

Part of this chapter was published in [Ric04].

Describing a sound in terms of morphological criteria requires to focus on the intrinsic perceptual qualities of sound. We saw that this listening mode is not natural and that we rather tend to 'hear' directly the sound-producing event or a message transmitted by a sound (see Chapter 2). In this chapter, we describe an experiment that aimed at investigating how easy it is for listeners with only minimal training to describe sounds according to our morphological sound description scheme. More specifically, we wanted to investigate how objective is this representation and how useful it could be as a search criterion in a sound retrieval system. We evaluated the usability of morphological representation for the retrieval of abstract and non-abstract sounds through an on line questionnaire, including sound examples and tests on the prototype application shown in section 4.5.

## 5.1 Material, procedure and subjects

After an introduction to morphological description (including sound examples) and to the prototype sound search engine, users were asked to label 10 sounds according to the dynamic profile (crescendo, decrescendo, iterative, impulse, delta, unvarying, other), the pitchness (pitched, complex, noisy), the pitchness profile (unvarying, varying) and the pitch profile (unvarying, varying-continuous, varying-stepped). This first part aimed at measuring the percentage of classification agreement over the subjects and to confirm our assumption that morphological description is listener-independent, or can be made so, providing a minimal amount of training.

In the second part, users had to listen to three abstract sounds and to retrieve them in a database of 100 abstract sounds using the prototype sound search engine. They could use traditional keywords and/or morphological classes. They were then asked whether they think that morphological representation was useful as search criteria for such sounds and if yes, whether it was useful as a main representation or as a complementary representation to traditional source-related keywords.

The third part consisted in the same test but, this time, focusing on non-abstract

sounds.

In the final part users were asked more general questions on the understandability of the current dimensions and their names and on the completeness of the description of each dimension (i.e. are there enough classes to describe each dimension?). Some comments were also gathered during informal discussions.

The questionnaire was answered by researchers in music technology, musicians and sound technicians. 14 people answered the listening test, and among them 10 answered the whole questionnaire. Most of the subjects (13) had medium or high musical training. Half of them used sound databases for professional purpose. Half of the subjects answered the questionnaire on line while the other half did it in our office.

## 5.2 Results

In order to measure how much listener-dependent was each dimension, we calculated a percentage of agreement, given by the number of sounds classified as the system did divided by the total number of answers (14 subjects*10 sounds) multiplied by 100. The results for the listening test are the following:

- 71% agreement for the dynamic profile type. Typical disagreements were between following classes: Delta / Crescendo or Decrescendo, Impulse / Decrescendo and Iterative / Any class. The two first disagreement types are due to the unclear boundaries between these classes. A sound having a Delta shaped amplitude envelope with a crescendo part much shorter than the decrescendo part could be perceived as globally decrescendo. Disagreements between the Iterative class and the other classes are all due to the fact that some people perceive the global envelope shape rather than each smaller entities a sound is made of (e.g. increasingly strong knocks on a door was often perceived as Crescendo, though, according to our definition, it is Iterative ).
- 78% agreement for the pitchness type. As expected, disagreements existed mainly between Pitched / Complex and Complex / Noisy classes. Once again disagreements are due to the unclear boundaries between the classes. Only one sound, a bouncing ping pong ball (sequence of very short impulses) was classified as both Noisy (11 subjects) and Pitched (3 subjects).
- Since pitch profile is (in our current system) only defined for pitched sounds, the result was calculated only for such sounds, which gives an agreement of 83%. Most disagreements happened in one sound made of two successive impulses having the same pitch (bike bell rings). Four subjects classified this sound as having a varying-stepped pitch profile, probably because of the iterative amplitude

envelope type.

- 73% agreement for pitchness profile. Despite the weakness of the model (see the notes on the evaluation of the model for this dimension in section 4.3), this result is not as bad as expected. As for the other dimensions, sounds having clearly varying or clearly unvarying pitchness profile were correctly classified with much higher agreement than for ambiguous sounds.
- All the subjects that answered the test on abstract sound retrieval (11 subjects) considered morphological labels as useful. 6 though that they should be use on their own (i.e. as a primary representation) and 5 though that they should be used in combination with traditional labels. Comments and suggestions included the possibility to describe the amount of pitched components vs. noisy components in sound having complex pitchness, the addition of a dynamic profile sub-category for specifying the attack (smooth, steep…) and the need for timbre description.

All the subjects that answered the test on non-abstract sound retrieval (10 subjects) considered morphological labels as useful in combination with traditional labels. In addition to those described above, more comments and suggestions were done, including the addition of more amplitude envelope types (without specifying what could be added) and some classes related to instrumental practice (e.g. glissando, pizzicato...).

The last part dealt with the understandability and the completeness of each dimension. All found that the *pitchness* is understandable and 3 found that it should be further described. One subject suggested distinguishing harmonic and inharmonic sounds in the pitched class. Dynamic profile is well understood by all subjects, but one suggested that the name 'amplitude envelope shape' would be more appropriate. Four subjects though that this dimension should be further described. Suggestions included adding tremolo and attack description as sub-dimensions, adding the possibility to combine the Iterative class with another one (e.g. Iterative-Crescendo) and adding a tool for drawing any envelope and retrieving corresponding sounds. Comments were done about the ambiguity between Delta and Crescendo or Decrescendo for some sounds. Pitchness profile was not well understood by 2 subjects. One pointed out that having three dimensions sharing the word *pitch* is misleading and the second commented that a sound could have two simultaneous unvarying and varying components. One subject also found that pitchness and pitchness profile were incompatible since if a sound is classified as noisy or pitched (this is not true for complex sounds), it should have an unvarying pitchness profile. Five subjects found that this dimension should be further described and suggested to add some typical profile, such as *pitched to noisy*. Pitch profile is not well understood by 2 subjects, one because of the use of the word *pitch* (see above) and the second because he did not understand well the class *Varying-Continuous*. Four subjects found that this dimension should be further described by adding some pitch contour classes

(ascending, descending…).

Surprisingly, only one subject suggested adding one morphological dimension, timbre (with no more details), to the current scheme.

No correlation was found between the results and the musical training or the use of sound databases for professional purpose.

## 5.3 Discussion

These results show that morphological labels are useful for retrieving abstract sounds as a primary or complementary representation and for non-abstract sounds as a complementary representation. Some useful comments were done in order to improve the system. The main problem seems to be that exclusive classes often lead to misclassification for ambiguous sounds. This could be solved by using fuzzy classification techniques, which give a probability of membership to all classes and allow then to discriminate typical sounds (e.g. 95% crescendo and 5% delta) from ambiguous sounds (e.g. 50% decrescendo, 40% impulse and 10% delta). It also seems sometimes difficult for users to perceptually separate the different morphological dimensions (see the bike bell rings example in results of the listening test for pitch profile). Since this way of listening (called *reduced listening* by Schaeffer) is unnatural, we assume that some more training would be sufficient to be able to focus on only one dimension. Some suggested features have been implemented (e.g. roughness and pitch contour) or will be considered for future work (e.g. analysis of each component of complex sounds). Adding a tool for drawing any profile is not planned yet because of the technical complexities it would amount.

# 6 Applications

In this chapter we review some potential applications of an automatic morphological description tool as well as some already existing systems.

## 6.1 Sound retrieval

Sound databases are used a lot by professionals of music composition and video or movie post production. Rapid development of information storage technology allows building larger and larger databases, making the retrieval of a specific sound a hard task. Typical commercial sound retrieval systems are usually source-centred, which means that retrieval is based on using the proper keywords or selecting the proper category that defines or specifies a sound source (e.g. [SFXLib]). In that context, sounds having no identifiable source (abstract sounds) can hardly be retrieved. Moreover, labels used for this type of sounds ("electronic", "weird", "FX"...) are often not consistent and vary from one company to another. Other approaches, based on perceptual description seem more suitable for abstract sound retrieval.

In the ECRINS project, a description scheme is defined, in which sounds are described according to the following dimensions, some of them being based on Schaeffer's morphology: Dynamic profile (amplitude evolution), melodic profile (pitch evolution), pitch, spectral distribution, and sound location [Geslin02]. Each of these dimensions is specified by a value or a set of values representing a typical case that can be used as a search criterion. Some of these values or sets of values are automatically estimated and the description can be refined manually by the user. However, the classification allowed by the automatic description is quite limited and the description scheme only includes pitched sounds. In Musclefish's system[25] , called Soundfisher, an automatic analysis is performed in order to classify or query sounds according their perceptual or acoustical content. The features computed include loudness, pitch, brightness, bandwidth (a measure of the spectrum width, its value is 0 for a pure tone and infinity for white noise), a measure related to timbre (MFCC) and their derivatives. Queries consist in specifying previously learned classes based on these features or in searching sounds similar to one provided by the user (search by similarity) by comparing the corresponding feature vectors [Wold96][26]. Another system, described in [McAd99], aimed at building a perceptual distance model from the physical attributes found to best correlate to the dimensions of the timbre space

25 www.**musclefish**.com/

26 Muscefish audio content retrieval technology has been recently integrated to Virage's AudioLogger (www.virage.com).

derived in [Krum89] and [McAd95] (see chapter 2 and 3). This distance was then used to retrieve sound by similarity in a database of harmonic musical sounds.

In the systems described above, the perceptual representation is either limited to harmonic sounds or only used for search by similarity. Harmonic sounds only lie in a small part of the much larger perceptual space we need to consider for abstract sounds, and it is not always possible to provide an example of a sound we want to retrieve. We think that computational morphological description, as described in chapter 4, could be of great interest for perceptual sound retrieval. The labels obtained could be used directly as search criteria, on their own or to refine a traditional search by source, and the LLDs used to estimate the labels can be used as features for a search by similarity, either global using the whole set of LLDs or by dimensions using specific LLDs (for example one could search for a sound that has a similar envelope to a given sound and a similar roughness to another one).

Our system was integrated in a prototype sound search engine combining semantic and perceptual representation to ease sound retrieval, described in [Cano04] (see section 4.5). A usability evaluation of morphological description for sound retrieval is described in Chapter 5.

## 6.2 Segmentation

Temporal audio segmentation is the process of dividing a sound stream into a sequence of elements. Algorithms are generally based on a two-steps process: first some short-term features, specific to the segmentation task to be performed (e.g. energy, phase, pitch, probability for a frame to belong to a class -e.g. speech or music -...) are computed and then a detection function (typically a detection of discontinuities) is computed from these features in order to estimate the boundaries of the segments.

Segmentation is the basis of music analysis or processing and most of the research on segmentation is done for such applications. In that context, the elements are notes or any other musical events. Typical algorithms for automatic segmentation of music consist in detecting the transients due to attacks at the beginning of the notes. Successful methods include detecting energy bursts in high frequency (e.g. [Masri96]) or in several frequency bands (e.g. [Klap99]), detecting discontinuities in the phase spectrum [Bello03] or combining amplitude and phase methods [Dux03]. A more complex system, using different features and detection functions for performing multi-level segmentation of musical signal (note segmentation, vibrato detection and speech/music discrimination) is described in [Ros00]. Another approach, that can be used for general-purpose segmentation, is to detect discontinuities in a measure of global distance between successive vectors of features representing different

dimensions of the sound (e.g. timbre, pitch, energy...) [Tzan99] [Foote00] [Zhang01].

All the algorithms described above are either specific to traditional musical sound or to limited classes (e.g. speech/music) or perform a somewhat arbitrary global segmentation of signals. By using morphological descriptors as the features, one could perform a segmentation according to any combination of the dimensions described in Chapter 2 (e.g. detection of noisy segments or detection of vibrato). This would be particularly useful for the analysis of electronic music, as discussed in 6.4.

## 6.3 Visualization

Sound visualization is the process of mapping sonic (physical or perceptual) parameters to visual parameters. A simple visual representation is the waveform, which is a plot of the amplitude variation against time, as shown in figure 6.1.

*Figure 6.1: Waveform of a fragment of a saxophone solo.*

Although the waveform does not provide much perceptual information on a sound, it is widely used for control or editing purpose. In sound editing applications, for instance, it allows controlling visually that a file actually contains a sound and eases sound handling tasks, such as selecting a segment for applying a given effect on it. Another visual representation of physical sonic parameters is the 3D representation of short-term frequency spectrum against time, called sonogram, shown in figure 6.2.

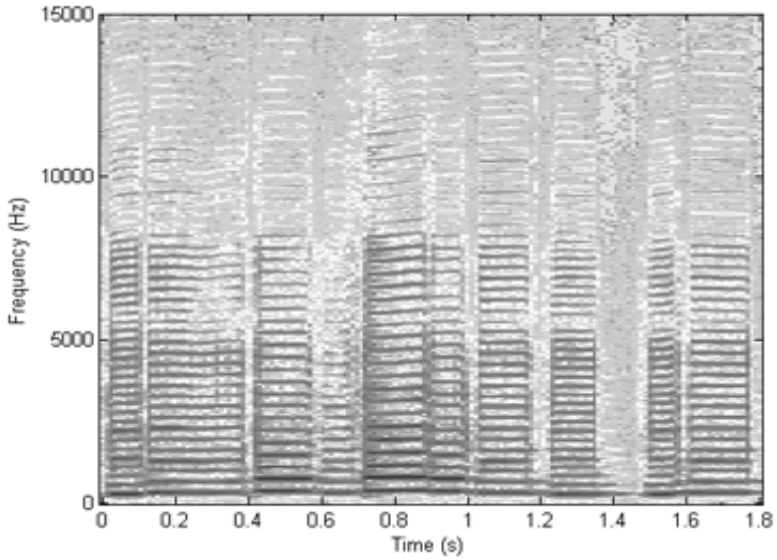



*Figure 6.2: Sonogram of the fragment of saxophone solo shown in figure 5.1.*

This representation contains a lot of data and is still not easily interpretable in terms of perceptual properties of a sound[27]. Some kind of visual summarization of the frequency axis into simple perceptual features have been proposed for sound retrieval purpose. Visual representations are of great interest for sound retrieval applications since some intuitive mapping allows visually browsing quickly a large amount of sounds. The commercial sound search engine designed by Comparisonics[28] shows for each sound the waveform coloured according to the frequency content. From the information given on the company's website, "shades of red are used for high-pitched sounds; greens and blues are used mostly for mid-range sounds; and bass sounds are represented by dark colors." An example of coloured waveform (in grayscale) is shown in figure 6.3.

27 The inverse process, i.e. the conversion of images in such time frequency representation, has been proposed as a sonic interpretation of the visual environment for blind people. A system based on this idea, as well as a description of the mapping used, is described at www.visualprosthesis.com and in [Meijer93]

28 http://www.findsounds.com/

*Figure 6.3: Waveform of a siren sound coloured according to the frequency content. Dark regions are high pitched and light regions are low-pitched. From www.findsounds.com*

In a similar representation, described in [Tzan02] and called timbregram, the waveform is coloured according to timbre features. One can see in these images a representation of the form, matter and variation criteria proposed by Schaeffer and described in 2.3.2. The form is given by the shape of the waveform, some matter criterion is represented by the colors and the variation of this criterion can be visually interpreted (e.g. in the siren sound shown in figure 6.3 the periodic frequency modulation is easily visible). Similar mappings using morphological criteria could allow users to choose the perceptual attribute they want to focus on, and switching between representations based on different mappings could ease and speed up the retrieval of a specific sound.

The visual representations described above are obtained by direct mapping of short term sound descriptors along a time axis. The perceptual identity of a sound is deduced by an interpretation of patterns of these descriptors (e.g. shape of the envelope, pitch trajectory...). Defining some perceptual categories, i.e. discrete values along each perceptual dimension, such as those proposed by Schaeffer (see 2.4), and mapping them to discrete values of visual shape, colour or texture, could provide symbolic representations for electronic music notation, as discussed in 6.4.

## 6.4 Electronic music notation

Musical representations are indispensable for analysis, interpreting or communicating a piece of music. The traditional western musical score, as shown in figure 2.3, has been used for a long time to represent music made of discrete notes from known instruments. Contemporary or electronic musics, which make use of a much larger range of sounds (e.g. noises, environmental sounds, sounds varying continuously...) cannot be represented using this notation, and composers of such music often devise their own notation, such as that shown in figure 6.4

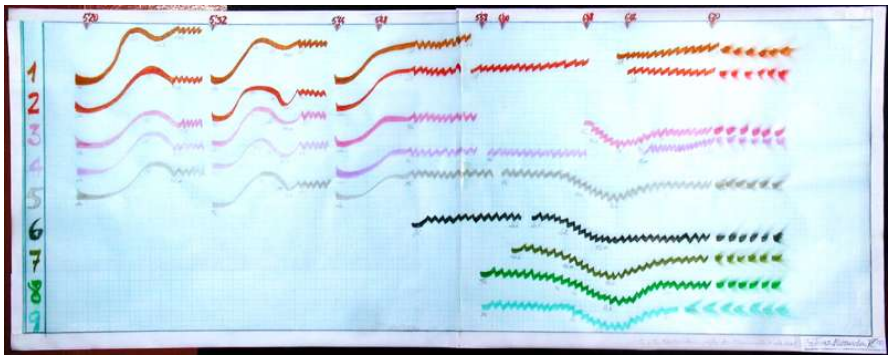


*Figure 6.4: Fragment of the score designed by Fátima Miranda for her piece "A Inciertas Edades" (from http://www.fatima-miranda.com).*

A program for the annotation of electronic music was designed by the 'Groupe de Recherche Musicale[29]' (Music Research Group) created by Pierre Schaeffer in the late 50's. This program, called Acousmographe, displays the waveform and a sonogram (see figure 6.3) that can be indexed and annotated by the user. The main feature is the possibility to manually add a graphical symbolic representation (coloured shape or imported image) to represent a segment selected by the user [Acous00]. An algorithm that retrieves segments perceptually similar, to speed up the annotation, will also be integrated [Spev02].

That was the primary objective of Schaeffer's typo-morphology (see 2.4) to provide a basis for the analysis of electronic music [Schaef66]. Automatic transcription systems for traditional western music (e.g. [Klap04]) performs automatic segmentation into notes and pitch estimation for each note. In the same way, morphological criteria could be used to automatically detect, describe and visualize sound objects in a piece of electronic music.

29 http://www.ina.fr/grm/

# 7 Conclusion and future work

In this chapter we summarize and discusse the work described in this document and propose some future work to be done for the final thesis.

## 7.1 Summary of contributions

We described in the first chapter of this document a study showing that listeners tend to talk about sounds by referring to the event that produced it. Recognising the source is the primary function of auditory perception and describing the sound itself is a difficult and unnatural task. We showed that describing a sound with no reference to its origin, or to a meaning it may contain, actually amounts to describe the perceptual image built by our auditory system from the acoustic stimulus. Auditory perception is far from being fully understood, but some experiments helped identify some dimensions of our auditory perceptual space. These studies often aimed at investigating specific dimensions from psycho-acoustical data (e.g. pitch, roughness...) or sub-spaces derived from subjective similarity judgments on pairs of specific sounds (e.g. musical or environmental sounds). The first step of our work consisted in reviewing the perceptual criteria found in these experiments (chapter 2) and the models that have been proposed for some of them (chapter 3). We showed that the perceptual space obtained by combining the results of these studies was similar to that defined in Pierre Schaeffer's work on sound objects description, so that his typo-morphology could serve as a good basis for a general perceptual representation. In chapter 4 we described a system that automatically generates a simplified perceptual description based on Schaeffer's morphological criteria. Some specific low-level features are extracted and directly assigned as numerical values for some criteria (e.g. roughness = '0.1') or combined by simple rules to identify a typical class for others (e.g. dynamic profile = 'impulsive', pitchness = 'noisy'...). An evaluation of the usability of this representation, described in chapter 5, showed that it is rather listener-independent and it is judged as useful in a sound retrieval task by a pool of users, both for abstract and non-abstract sounds. Finally, we described in chapter 6 some potential applications of computational morphological description and showed that it would be of great interest for segmenting, retrieving or visualising abstract sounds.

## 7.2 Discussion and future work

The work reported in this document suggests that a perceptual representation is needed to handle abstract sounds in content-based audio description systems. It also shows that perceptual criteria can be automatically extracted by some algorithms based on current knowledge on auditory perception. Further research will consist in completing the description scheme, improving the computational models used and investigating some potential applications described in Chapter 5.

### 7.2.1 Completing the description scheme

Our current description scheme, shown in table 4.1, has been built progressively from Schaeffer's typo-morphology (see section 2.4.3 or Appendix 1), starting by the criteria that were found to be salient in our perception of environmental sounds, which exhibits a complete range of the qualities we aim at describing. The resulting representation is not complete and, although we still don't know what a complete perceptual representation would be, it seems clear that some important features are missing. Remaining criteria from Schaeffer's typo-morphology should first be considered, e.g. characterization of the attack, description of the velocity and periodicity of the pitch profile or detection and characterization of vibrato and tremolo, as well as further description of what Schaeffer called *harmonic timbre* (e.g. 'metallicness', 'richness'...) and further characterization of varying pitchness profile (e.g. 'noisy to pitched'...). Other important missing features are the loudness and the characterization of the frequency content of noisy sounds, the 'pitch of noise', to discriminate, for instance, low, middle and high frequency noises. Table 2.1, though not exhaustive, will provide a basis for future work.

### 7.2.2 Improving the computational models, investigating new features and testing fuzzy classification techniques

The computational models used in our system are quite simple and appeared to perform poorly for certain sounds. The onset detection is based on energy discontinuities and failed at detecting smooth transients, such as those played by bowed instruments or singing voice. More complex algorithms, such as that combining magnitude and phase discontinuities detection (see section 6.2) should be investigated. Regarding pitch estimation, the model we used performs well for monophonic harmonic sounds but does not explicitly detect several pitched components. Multi-pitch analysis would allow detecting and further describing polyphonic sounds[30] (e.g. number of components, pitches and magnitudes) and would

30 An algorithm for multi-pitch estimation, only evaluated for musical sounds, is described by Klapuri

therefore improve pitchness and pitchness profile classification. The simplified model of roughness we tested recently allows discriminating extreme sounds (very rough vs. very smooth) with correct performance, but the complete model should definitely be implemented for a better estimation of a whole range of roughness. Some methods for estimating the new criteria proposed in 7.2.1 (e.g. metallicness´ or 'pitch of noise') should also be investigated.

The features extracted from the amplitude envelope for dynamic profile classification performs well and are not planned to be modified. Regarding pitch profile classification, the main problem seems to be the pitch estimation itself rather than the set of features, which are similar to those used for dynamic profile classification. In the case of pitchness classification, only one feature is used, the pitch salience, which detects with good performance only extreme sounds (noise or monophonic harmonic sounds) while complex sounds are detected somewhat by default, i.e. as not being a noise nor a pitched sounds. Features extracted from multi-pitch estimation, e.g. noisy to pitched components energy ratio and number of components, are likely to perform much better.

Finally, while classification trees perform well when applied to sounds clearly belonging to a given class, it can lead to misclassification for ambiguous sounds, lying at the boundary between two classes. Some fuzzy classification techniques should be more appropriate and will be investigated.

### 7.2.3 Investigating potential applications

The main practical goal of designing a perceptual audio description scheme is to provide abstract sounds with a representation to handle them in the content-based applications described in Chapter 6, including parametric perceptual segmentation, retrieval and visualization. This representation could also be used as a complement of causal and musical representations for non-abstract sounds, as proposed in Chapter 5. As a final objective, we would like to combine these applications to investigate the use of morphological description for electronic music notation.

in [Klap04].

# Bibliography


[Acous00] *Acousmographe*. Technical Documentation, 2000

[Appell02] J. E. Appell. *Loudness Models for Rehabilitative Audiology*. Dissertation. Oldenburg University, 2002.

[Bello03] J. P. Bello and M. Sandler. *Phase-Based Note Onset Detection for Music Signals*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2003.

[Berger64] K. W. Berger. *Some Factors in the Recognition of Timbre*. Journal of the Acoustical Society of America, 36, 1964.

[Bism74] G. von Bismarck. *Timber of Steady Sounds: A Factorial Investigation of its Verbal Attributes*. Acustica, 30, 1974.

[Bjork85] E. A. Bjork. *The Perceived Quality of Natural Sounds*. Acustica, 57, 1985.

[Bone01] T. L. Bonebright. *Perceptual Structure of Everyday Sounds: A Multidimensional Scaling Approach*. Proceedings of the 2001 International Conference on Auditory Display, 2001.

[Bregman90] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[Brown91] J. C. Brown and B. Zhang. *Musical Frequency Tracking Using the Method of Conventional and "Narrowed" Autocorrelations*. Journal of the Acoustical Society of America, 89(5), 1991.

[Brown92] J. C. Brown. *Musical Fundamental Frequency Tracking Using a Pattern Recognition Method*. Journal of the Acoustical Society of America, 92(3), 1992.

[Cabrera99] D. Cabrera. *PsySound: A Computer Program for Psychoacoustical Analysis*. Proceedings of the Australian Acoustical Society Conference, 1999.

[Cano04] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard and N. Wack. *Knowledge and Perceptual sound Effects Asset Management*. Proceedings of the 1st International Conference on E-Business and Telecommunication Networks, 2004.

[Cano98] P. Cano. *Fundamental Frequency Estimation in the SMS Analysis*. Proceedings of COST G6 Conference in Digital Audio Effects, 1998.

[Casey98] M. Casey. *Auditory Group Theory with Applications to Statistical Basis Method for Structured Audio*. PhD Thesis. Massachusetts Institute of Technology, 1998.

[Celma04] O. Celma, E. Gomez, J. Janer, F. Gouyon, P. Herrera and D. Garcia. *Tools for Content-Based Retrieval and Transformation of Audio Using MPEG-7: The SPOffline and the MDTools*. Proceedings of the 25th International Audio Engineering Society Conference, 2004.

[Chion83] M. Chion. *Guide des Objets Sonores*. INA-GRM, Buchet/Chastel, 1983.

[Dan02] R. B. Dannenberg and Ning Hu. P*attern Discovery Techniques for Music Audio*. Proceedings of the 3rd International Conference on Music Information Retrieval, 2003.

[Donnadieu94] S. Donnadieu, S. McAdams and S. Winsberg. *Context Effects in Timbre Space*. Proceedings of the 3rd International Conference on Music Perception and Cognition, 1994.

[Dov93] B. Doval and X. Rodet. *Fundamental Frequency Estimation and Tracking Using Maximum Likelihood Harmonic Matching and HMM's*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1993.

[Dux03] C. Duxbury, J. P. Bello, M. Davies and M. Sandler. *Complex Domain Onset Detection for Musical Signals*. Proceedings of the 6th International Conference on Digital Audio Effects, 2003.

[Faure00] A. Faure. *Des Sons aux Mots, Comment Parle-t-on du Timbre Musical?*. PhD Thesis. Ecole des Hautes Etudes en Sciences Sociales, 2000.

[Fletcher33] H. Fletcher and W. Munson. *Loudness, its Definition, Measurement and Calculation*. Journal of the Acoustical Society of America, vol. 5, 1933.

[Foote00] J. Foote. *Automatic Audio Segmentation Using a Measure of Audio Novelty*. Proceedings of the IEEE International Conference on Multimedia and Expo, 1, 2000.

[Gaver93] W. W. Gaver. *What in the World Do We Hear?*. Ecological Psychology, 5 (1), 1993.

[Geslin02] Y. Geslin, P. Mullon and M. Jacob. *Ecrins: An Audio Content Description Environment for Sound Samples*. Proceedings of the International computer Music Conference, 2002.

[Gibson66] J. J. Gibson. *The Senses Considered as Perceptual Systems*. Houghton Mifflin, 1966.

[Gomez03a] E. Gomez, A. Klapuri and B. Meudic. Melody Description in the Context of Music Content Processing. Journal of New Music Research, 32(1), 2003.

[Gomez03b] E. Gomez, F. Gouyon, P. Herrera and X. Amatriain. *Using and Enhancing the Current MPEG-7 Standard for a Music Content Processing Tool*. Proceedings of the 114th Convention of the Audio Engineering Society, 2003.

[Gouyon03] F. Gouyon and B. Meudic. *Towards Rhythmic Content Processing of Musical Signals: Fostering Complementary Approach*. Journal of New Music Research, 32(1), 2003.

[Green01] S. Greensberg and M. Slaney (Editors). *Computational Models of Auditory Function*. IOS Press, 2001.

[Grey77] J. M. Grey. *Multidimensional Perceptual Scaling of Musical Timbre*s. Journal of the Acoustical Society of America, 61(5), 1977.

[Gygi00] B. Gygi, G. R. Kidd and C.S. Watson. *Identification and Similarity Judgments of Environmental Sounds*. Journal of the Acoustical Society of America, 107(2), 2000.

[Gygi01] B. Gygi. *Factors in the Identification of Environmental Sounds*. PhD Thesis. Indiana University, 2001.

[Handel04] S. Handel and M. L. Erickson. *Sound Source Identification: The Possible Role of Timbre Transformations*. Music Perception, 21(4), 2004.

[Handel89] S. Handel. *Listening: An Introduction to the Perception of Auditory Events*. MIT Press, 1989.

[Helm54] H. von Helmholtz. *On the Sensations of Tone*. Dover Publications, 1954.

[Herrera03] P. Herrera, G. Peeters and S. Dubnov. *Automatic Classification of Musical Instrument Sounds*. Journal of New Music Research, 32, 2003.

[Howard77] J. H. Howard. *Psychophysical Structure of Eight Complex Underwater Sounds*. Journal of the Acoustical Society of America, 62(1), 1977.

[Kabal02] P. Kabal *An Examination and interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality*. Technical report, 2002.

[Klap04] A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD Thesis. Tampere University of Technology, 2004.

[Klap99] A. Klapuri. *Sound Onset Detection by Applying Psychoacoustic Knowledge*. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1999.

[Krim94] J. Krimphoff, S. McAdams and S. Winsberg. *Caractérisation du Timbre des Sons Complexes. II: Analyses Acoustiques et Quantification Psychophysique.* Journal de Physique, 4, 1994.

[Krum89] C. L. Krumhansl. *Why is Musical Timbre So Hard to Understand?*. In Structure and Perception of Electroacoustic Sound and Music. Editors: S. Nielzenand and O. Olsson, Elsevier, 1989.

[Kun96] N. Kunieda, T. Shimamura and J. Susuki. *Robust Method of Measurement of Fundamental Frequency by ACLOS - Autocorrelation of Log Spectrum.* Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1996.

[Lahat87] M. Lahat, R. j: Niederjohn and D .A. Krubsack. *A Spectral Autocorrelation Method for Measurement of the Fundamental Frequency of Noise-Corrupted Speech*. IEEE Transactions on Acoustics, Speech and Signal Processing, 35(6), 1987.

[Laka00] S. Lakatos. *A Common Perceptual Space for Harmonic and Percussive Timbres*. Perception and Psychophysics, 62, 2000.

[Macph95] E. A. Macpherson. *A Review of Auditory Perceptual Theories and the Prospects for an Ecological Account.* Dissertation, 1995.

[Maher94] R. C. Maher and J. W. Beauchamp. *Fundamental Frequency Estimation of Musical Signals Using a Two-Way Mismatch Procedure.* Journal of the Acoustical Society of America, 95(4), 1994.

[Mar03] J. M. Martinez. *MPEG-7 Overview*. http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[Martin96] K. D. Martin. *A Blackboard System for Automatic Transcription of Simple Polyphonic Music.* Massachusetts Institute of Technology, Media Laboratory, Perecptual Computing Section. Technical report No. 385, 1996.

[Masri96] P. Masri and A. Bateman. *Improved Modelling of Attack Transients in Music Analysis-Resynthesis*. Proceedings of the International Computer Music Conference, 1996.

[McAd93] S. McAdams and E. Bigand. *Thinking in Sound: The Cognitive Psychology of Human Audition*. Oxford University Press, 1993.

[McAd95] S. McAdams, S. Winsberg, G. de Soete and J. Krimphoff. *Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities and Latent Subject Classes*. Psychological Research, 58, 1995.

[McAd99] S. McAdams and N. Misdariis. *Perceptual-Based Retrieval in Large Musical Sound Databases*. Proceedings of the Human Centered Processes Conference, 1999.

[Meddis97] R. Meddis and L. O'Mard. *A Unitary Model of Pitch Perception*. Journal of the Acoustical Society of America, 102(3), 1997.

[Meijer93] P. B. L. Meijer. *An Experimental System for Auditory Image Representations*. IEEE Transactions on Biomedical Engineering, 39(2), 1993.

[Moore83] B. C. J. Moore and B. R. Glasberg. *Suggested Formulae for*

*Calculating Auditory Filter Bandwidths and Excitation Patterns*. Journal of the Acoustical Society of America, 74, 1983.

[Moore96] B. C. J. Moore and B. R. Glasberg. *A Revision of Zwicker's Loudness Model*. ACTA Acustica, 82, 1996.

[Moore97] B. C. J. Moore and B. R. Glasberg. *A Model for the Prediction of Thresholds, Loudness and Partial Loudness*. Journal of the Engineering Society, 45, 1997.

[Noll67] A. M. Noll. *Cepstrum Pitch detection*. Journal of the Acoustical Society of America, 41(2), 1967.

[Ong04] B. Ong and P. Herrera. *Computing Structural Descriptions of Music through the Identification of Representative Excerpts from Audio Files*. Proceedings of the 25th International AES Conference, 2004.

[Peeters00] G. Peeters, S. McAdams and P. Herrera. *Instrument Description in the Context of MPEG-7*. Proceedings of International Computer Music Conference, 2000.

[Press98] D. Pressnitzer. *Perception de Rugosité Psychoacoustique: D'un Attribut Elementaire de l'Audition a l'Ecoute Musicale*. PhD Thesis. Université Paris 6. 1998

[Ric03a] J. Ricard and P. Herrera. *Towards Computational Morphological Sound Description*. Proceedings of IRCAM MusicNetwork Workshop, 2003.

[Ric03b] J. Ricard and P. Herrera. *Using Morphological Sound Description for Generic Sound Retrieval*. Proceedings of the 4th International conference on Music Information Retrieval, 2003.

[Ric04] J. Ricard and P. Herrera. *Morphological Sound Description: Computational Model and Usability Evaluation*. Proceedings of the 116th Convention of the Audio Engineering Society, 2004.

[Rocch03] D. Rocchesso and F. Fontana (Editors). *The Sounding Object*. 2003. Available on line at http://www.soundobject.org/

[Ros00] S. Rossignol. *Segmentation et Indexation des Signaux Sonores Musicaux.* PhD Thesis. Université Paris 6, 2000.

[Schaef66] P. Schaeffer. *Traité des Objets Musicaux*. Seuil, 1966.

[Schafer77] R. M. Schafer. *The Soundscape*. Destiny Books, 1977.

[Serra98] X. Serra and J. Bonada. *Sound Transformations Based on the SMS High Level Attributes*. Proceedings of COST G6 Conference on Digital Audio Effects, 1998.

[SFXLib] *The Sound Effects Library*. http://www.sound-effects-library.com/.

[Slaney98] M. Slaney. *Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work*. Technical Report, 1998

[Spev02] C. Spevak and E. Favreau. *Soundspotter - A Prototype System for Content-Based Audio Retrieval*. Proceedings of the 5th International Conference on Digital Audio Effects, 2002.

[Stevens37] S. S. Stevens, J. Volkmann and E. B. Newman. *A Scale for the Measurement of the Psychological Magnitude Pitch*. Journal of the Acoustical Society of America, 8, 1937.

[Talkin95] D. Talkin. *A Robust Algorithm for Pitch Tracking*. In Kleijn and Paliwal, editors, Speech Coding and Synthesis,1995

[Tzan02] G. Tzanetakis. *Manipulation, Analysis and Retrieval Systems for Audio Signals*. PhD Thesis. Princeton University, 2002.

[Tzan99] G. Tzanetakis and P. Cook. *Multi-Feature Audio Segmentation for Browsing and Annotation*. Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1999.

[Vander79] N. J. Vanderveer. *Ecological Acoustics: Human Perception of Environmental Sounds*. Dissertation Abstracts International. 40/09B, 4543. University Microfilms No. 8004002, 1979.

[Wessel79] D. Wessel. *Timbre Space as a Musical Control Structure*. Computer Music Journal, 3(2), 1979.

[Win92] P. H. Winston. *Artificial Intelligence.* Addison-Wesley, 1992.

[Wold96] E. Wold, T. Blum, D. Keislar and J. Wheaton. *Content-Based Classification, Search and Retrieval of Audio*. IEEE Transactions on Multimedia, 3 (2), 1996.

[Yost00] W. A. Yost. *Fundamentals of Hearing: An Introduction*. Academic Press, 2000.

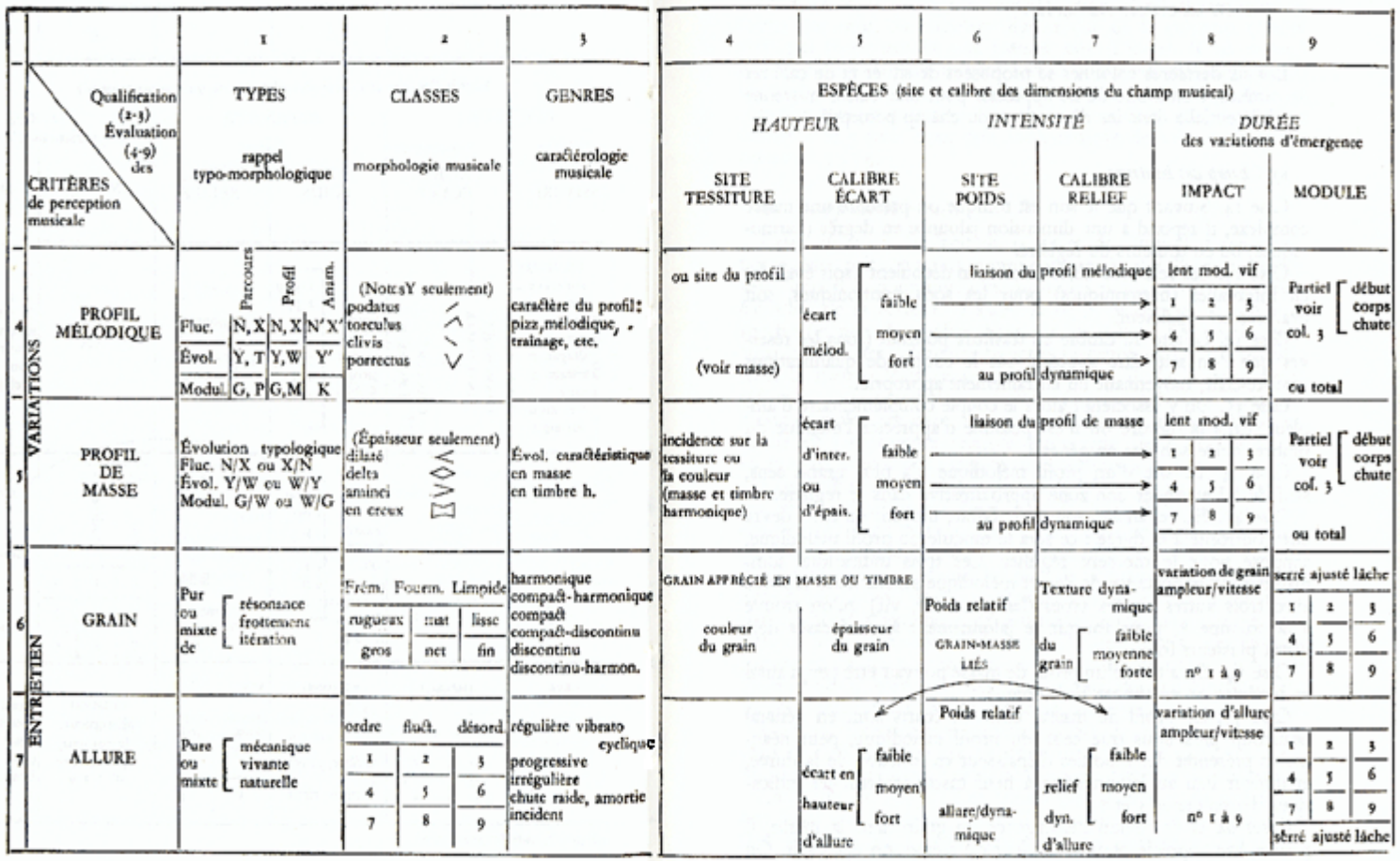
[Zhang01] T. Zhang and C. C. J. Kuo. *Audio Content Analysis for Online Audiovisual Data Segmentation and Classification*. IEEE Transactions on Speech and Audio Processing, 2001.

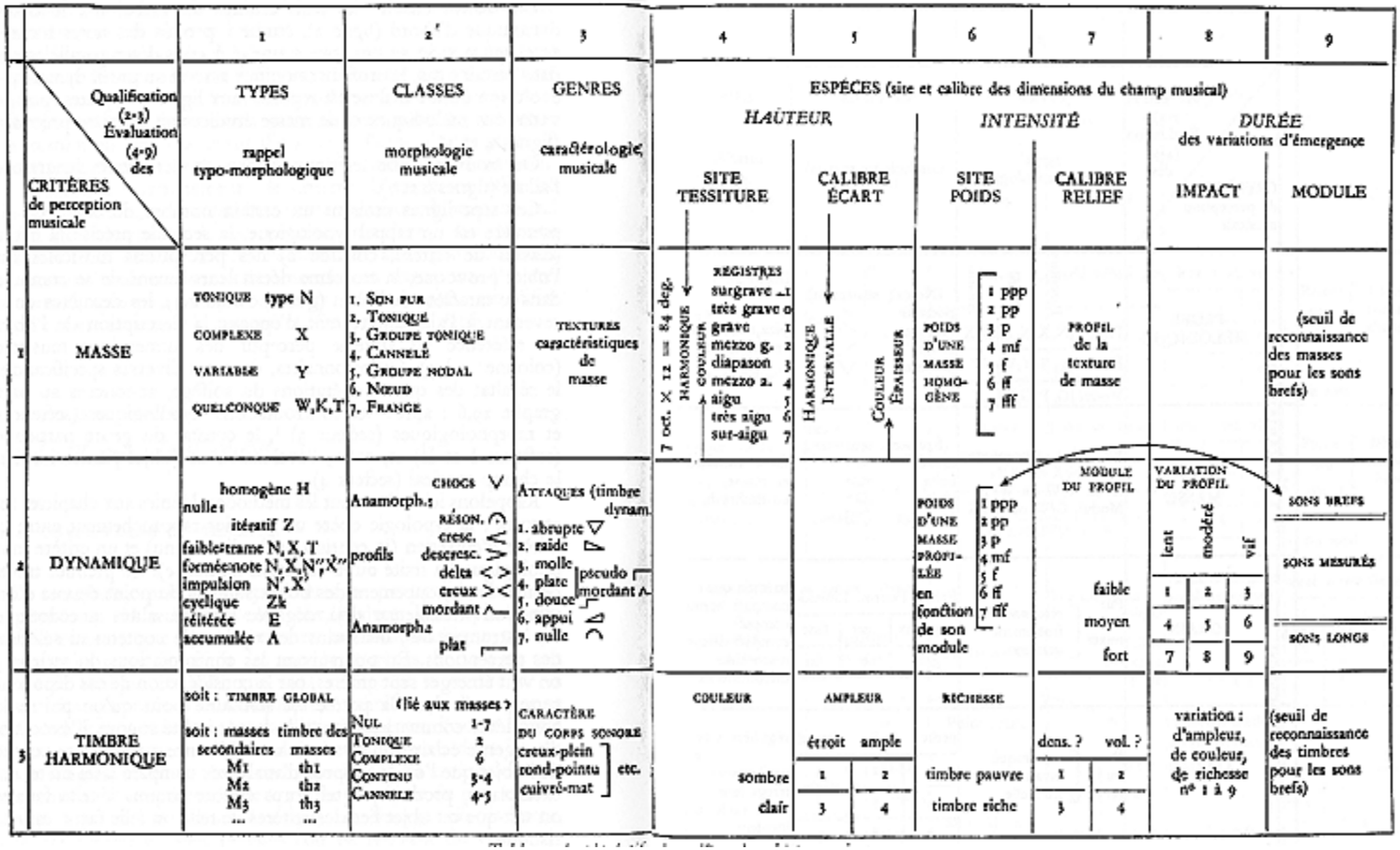
[Zhang99] T. Zhang and C.-C. Jay Kuo. *Classification and Retrieval of Sound Effects in Audiovisual Data Management*. Proceedings of 33rd Asilomar Conference on Signals, Systems and Computers, 1999.

[Zwicker90] E. Zwicker and H. Fastl. *Psycho-Acoustics: Facts and Models*. Springer, 1990.

# Appendix 1

# Schaeffer's typo-morphology summary table


1
2
3
4
5
6
7
8
9
Qualification (2-3) Évaluation (4-9) des
CRITÈRES de perception musicale
TYPES
rappel typo-morphologique
CLASSES
morphologie musicale
GENRES
caractérologie musicale
VARIATIONS
ENTRETIEN
PROFIL MÉLODIQUE
Parcours
Profil
Anim.
Fluc.
N, X
N, X
N' X'
Évol.
Y, T
Y, W
Y'
Modul.
G, P
G, M
K
(Notes Y seulement)
podatus
torculus
clivis
porrectus
caractère du profil: pizz, mélodique, trainage, etc.
PROFIL DE MASSE
Évolution typologique
Fluc. N/X ou X/N
Évol. Y/W ou W/Y
Modul. G/W ou W/G
(Épaisseur seulement)
dilaté
delta
aminci
en creux
Évol. caractéristique en masse en timbre h.
GRAIN
Pur ou mixte de
résonance
frottement
itération
Frém. Fourn. Limpide
rugueux
mat
lisse
gros
net
fin
harmonique
compact-harmonique
compact
compact-discontinu
discontinu
discontinu-harmon.
ALLURE
Pure ou mixte
mécanique
vivante
naturelle
ordre
fluct.
désord.
régulière vibrato
cyclique
progressive
irrégulière
chute raide, amortie
incident
4
5
6
7
8
9
ESPÈCES (site et calibre des dimensions du champ musical)
HAUTEUR
INTENSITÉ
DURÉE des variations d'émergence
SITE TESSITURE
CALIBRE ÉCART
SITE POIDS
CALIBRE RELIEF
IMPACT
MODULE
ou site du profil
(voir masse)
écart mélod.
faible
moyen
fort
liaison du
profil mélodique
au profil
dynamique
lent mod. vif
Partiel voir col. 3
début
corps
chute
ou total
incidence sur la tessiture ou la couleur (masse et timbre harmonique)
écart d'inter. ou d'épais.
profil de masse
GRAIN APPRÉCIÉ EN MASSE OU TIMBRE
couleur du grain
épaisseur du grain
Poids relatif
GRAIN-MASSE LIÉS
Texture dynamique du grain
faible
moyenne
forte
variation de grain ampleur/vitesse
nº 1 à 9
serré ajusté lâche
Poids relatif
écart en hauteur d'allure
relief dyn. d'allure
allure/dynamique
variation d'allure ampleur/vitesse


1
2'
3
4
5
6
7
8
9
Qualification (2-3) Évaluation (4-9) des
CRITÈRES de perception musicale
TYPES
rappel typo-morphologique
CLASSES
morphologie musicale
GENRES
caractérologie musicale
ESPÈCES (site et calibre des dimensions du champ musical)
HAUTEUR
INTENSITÉ
DURÉE des variations d'émergence
SITE TESSITURE
CALIBRE ÉCART
SITE POIDS
CALIBRE RELIEF
IMPACT
MODULE
1 MASSE
TONIQUE type N
COMPLEXE X
VARIABLE Y
QUELCONQUE W,K,T
1. Son pur
2. Tonique
3. Groupe tonique
4. Cannelé
5. Groupe nodal
6. Nœud
7. Frange
TEXTURES caractéristiques de masse
7 oct. × 12 = 84 deg.
HARMONIQUE
COULEUR
REGISTRES
surgrave -1
très grave 0
grave 1
mezzo g. 2
diapason 3
mezzo a. 4
aigu 5
très aigu 6
sur-aigu 7
HARMONIQUE
INTERVALLE
COULEUR
ÉPAISSEUR
POIDS D'UNE MASSE HOMOGÈNE
1 ppp
2 pp
3 p
4 mf
5 f
6 ff
7 fff
PROFIL de la texture de masse
(seuil de reconnaissance des masses pour les sons brefs)
2 DYNAMIQUE
homogène H
nulle :
itératif Z
faible trame N, X, T
formée note N, X, N', X''
impulsion N', X'
cyclique Zk
réitérée E
accumulée A
CHOCS
Anamorph. :
RÉSON.
cresc.
profils descresc.
delta
creux
mordant
Anamorph. :
plat
Attaques (timbre dynam.
1. abrupte
2. raide
3. molle
4. plate
5. douce
6. appui
7. nulle
pseudo mordant
POIDS D'UNE MASSE PROFILÉE en fonction de son module
1 ppp
2 pp
3 p
4 mf
5 f
6 ff
7 fff
MODULE DU PROFIL
faible
moyen
fort
VARIATION DU PROFIL
lent
modéré
vif
1 2 3
4 5 6
7 8 9
SONS BREFS
SONS MESURÉS
SONS LONGS
3 TIMBRE HARMONIQUE
soit : TIMBRE GLOBAL
soit : masses secondaires timbre des masses
M1 th1
M2 th2
M3 th3
(lié aux masses)
Nul 1-7
Tonique 2
Complexe 6
Contenu 5-4
Cannelé 4-5
CARACTÈRE DU CORPS SONORE
creux-plein
rond-pointu
cuivré-mat
etc.
COULEUR
sombre
clair
AMPLEUR
étroit ample
1 2
3 4
RICHESSE
timbre pauvre
timbre riche
dens. ? vol. ?
1 2
3 4
variation : d'ampleur, de couleur, de richesse n° 1 à 9
(seuil de reconnaissance des timbres pour les sons brefs)

# Appendix 2 Low-level descriptors used for morphological classification

### Dynamic profile

All LLDs were computed on the log-amplitude envelope A(n), of size N.

- Flatness coefficient: ratio of the value above which lie 5% of the values to the value above which lie 80% of the values:

$$Flatness = \frac{E_5}{E_{80}}$$

where $E_X$ is the value above which lie X% of A(n), given by

$$E_X = A_{sorted}\left(round\left(N\frac{\left(100-X\right)}{100}\right)\right)$$

This coefficient is close to one for flat envelope and large for sounds having a large dynamic.

- Number of onsets, detected by looking for peaks above a threshold on the amplitude envelope derivative. If the number of onsets is greater than 1 (the first onset correspond to the attack of the sound), the sound is classified as iterative.

- Maximum amplitude time to total length ratio, given by

$$MaxToTot = \frac{n_{max}}{N}$$

where $A\left(n_{max}\right)=max\left(A\left(n\right)\right)$

This coefficient show how much the maximum amplitude is off-center. Its value is close to 0 if the maximum is close to the beginning (e.g. Decrescendo or Impulsive sounds), close to 0.5 if it is close to the middle (e.g. Delta sounds) and close to 1 if it is close to the end of the sound (e.g. Crescendo sounds).

- Temporal centroid to total length ratio, given by

$$TCToTot=\frac{TC}{N}$$

where the temporal centroid TC is given by

$$TC=\frac{\sum_{n=0}^{n=N-1} A\left(n\right)\cdot n}{\sum_{n=0}^{n=N-1} n}$$

This coefficient show how the sound is 'balanced'. It is close to 0 if most of the energy lies at the beginning (e.g. Decrescendo or Impulsive sounds), close to 0.5 is the sound if symmetric (e.g. Unvarying or Delta sounds) and close to 1 if most of the energy lies at the end of the sound (e.g. Crescendo sounds).

- Derivative average, weighted by the amplitude, after the maximum amplitude, given by

$$DerAvAfterMax=\frac{\sum_{k=n_{max}}^{k=N-1}\left(A\left(n\right)-A\left(n-1\right)\right)\times A\left(n\right)}{\sum_{k=n_{max}}^{k=N-1} A\left(n\right)}$$

This coefficient helps discriminating Impulsive sounds, which have a steepest release, so a smaller value, from Decrescendo sounds.

- Maximum derivative before the maximum, given by

$$MaxDerBeforeMax = max\left( A\left(n\right) - A\left(n-1\right)\right),\ for\ n \in \left[1 \quad n_{max}\right]$$

This coefficient helps discriminating Crescendo and Delta sounds, that have a smooth attack, so a smaller value than sounds with different dynamic profile.

**Pitchness**

- Average of the short-term pitch salience weighted by the short-term energy. The pitch salience is given by the ratio of the highest peak to the 0-lag peak in the autocorrelation function (ACF).

The ACF r(n) is given for a discrete signal x(n) by

$$r\left(n\right) = \frac{1}{K} \sum_{k=0}^{K-n-1} x\left(k\right) \times x\left(k+n\right)$$

and the pitch salience is given by

$$Pitch\ salience = \frac{max\left(r\left(n\right)\right)\ for\ n \in \left[n_{min} \quad n_{max}\right]}{r\left(0\right)}$$

where $n_{min}$ and $n_{max}$ are specified by the pitch analysis range, as explained in 4.2.1.

Unpitched sounds have a value close to 0 while harmonic sounds have a value close to 1.

**Pitchness profile**

- Pitch salience variance

The variance of a function f(p) of length P is given by

$$Var\left(f\left(p\right)\right)=\frac{1}{P}\sum_{p=0}^{P-1}\left(f\left(p\right)-\mu\right)^2$$

where μ is the average value of f(p).

Sounds having Unvarying pitchness have a small value while sounds having varying pitchness have a high value.

**Pitch profile**

All LLDs were computed on the log-pitch envelope Pitch(n). Some segmentation was performed by detecting peaks above a threshold in the pitch envelope acceleration. The following LLDs were then used:

For Unvarying/Varying-Continuous/Varying-stepped pitch profile classification:

- Pitch(n) variance.

  Sounds having Unvarying pitch have a small value while sounds having Varying pitch have a high value.

- Average of the log pitch variances computed in each segment.

  Sounds having a pitch varying by steps have small value (for such sounds, one segment correspond to one step, in which the pitch does not vary) while sounds having a pitch varying continuously have a high value (in that case, no discontinuities are detected, so the variance is computed over the whole sound).

For Delta/Inverse delta/Ascending/Descending/Other pitch profile classification (test):

- Minimum pitch time to total length ratio.

  Sounds having an ascending pitch a value close to 0.

- Pitch centroid (center of gravity of the pitch).

  This LLD is similar to the temporal centroid use for dynamic profile classification.

- Ratio of energy after the maximum to energy before the maximum.

  Sounds having an ascending pitch have a small while sounds having a descending pitch have a high value.

# Appendix 3 Publications related to the project

[Cano04] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard and N. Wack. Knowledge and Perceptual sound Effects Asset Management. Proceedings of the 1st International Conference on E-Business and Telecommunication Networks, 2004.

[Ric03a] J. Ricard and P. Herrera. Towards Computational Morphological Sound Description. Proceedings of IRCAM MusicNetwork Workshop, 2003.

[Ric03b] J. Ricard and P. Herrera. Using Morphological Sound Description for Generic Sound Retrieval. Proceedings of the 4th International conference on Music Information Retrieval, 2003.

[Ric04] J. Ricard and P. Herrera. Morphological Sound Description: Computational Model and Usability Evaluation. Proceedings of the 116th Convention of the Audio Engineering Society, 2004.