A proposal for the description of audio in the context of MPEG-7

Perfecto Herrera, Xavier Serra

Audiovisual Institute – Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
http://www.iua.upf.es
{pherrera, xserra}@iua.upf.es

Geoffroy Peeters

IRCAM
1, Place Igor Stravinsky, 75004 Paris, France
http://www.ircam.fr
Geoffory.Peeters@ircam.fr

[Published in the Proceedings of the CBMI'99 European Workshop on Content-Based Multimedia Indexing]

Abstract

Sound content description is one of the aims of the MPEG-7 initiative. Although MPEG-7 focuses on indexing and retrieval of audio, there are other sound content-based processing applications waiting to be developed once we have a robust set of descriptors and structures for putting them into relation, and for expressing semantic concerns about sound. Spectral Modeling techniques provide one usable framework for extracting and organizing sound content descriptions. In this paper we will introduce one particular approach to spectral modeling, then we will present some sound descriptors that can be derived from them in order to develop sound descriptions, and we will discuss the features of a structure for organizing the information that can be derived from them (a so called "Description Scheme"). All of our current descriptors can be considered low- or mid-level, thus we will not cover the high level description of music (musical forms and styles, roles of characters in a movie, etc.) which is also relevant in MPEG-7 indeed. The descriptors proposed are the result of a sound analysis based on a spectral modeling technique, and for all of them we have devised automatic extraction procedures. The Description Scheme we present is intended to be a generic one that, based on a hierarchical (and recursive in some places) structure, can describe sound at multiple levels of detail, addressing both syntactic (structural) and semantic (content) ways for describing sound.

1. Introduction

MPEG-7 is an standardization initiative of the Motion Pictures Expert Group that, instead of focusing on audio coding like MPEG-1, MPEG-2 and MPEG-4, is meant to be an standardization of the way to describe sound [1]. The main application for MPEG-7 should be the content-based indexing and retrieval of audio, or of any other media. Although content based audio descriptions can prove extremely fruitful in conventional tasks as audio editing, music composing, sound effects selection, or video cueing, they can also open new possibilities for live music mixing and DJing, sound signaturing for copyright protection, music commercial assessment and recommendation, agent based TV scheduling, etc.

Audio content extraction and managing can be achieved by different, sometimes overlapping, and anyway non-exclusive means: there are traditional signal processing techniques, computational auditory scene analysis techniques, statistical techniques, etc. [2, 3, 4, 5, 6]. There are also manual keyword labeling techniques and thesauri that can work synergistically with automatic techniques in order to help to describe and organize more thoroughly the content of audio material. The key issue here is that MPEG-7 will not standardize the way to obtain these descriptions nor how to use them, but only the descriptions and the way for structuring them.

Describing sound content involves using procedures, techniques, and data, that have been found and developed in different research areas (i.e. signal processing, music cognition, artificial intelligence, etc.), in order to solve problems as *Sound segregation* (components of sound mixes need to be identified in order to describe them separately [7]), *Segmentation* (time-localized abrupt changes in significative parameters of sound have to be detected and classified as diagnostic cues for understanding changes in the content of the sonic flow), *Sound event and source characterization* (detecting pitch notes and durations, chords, expressive gestures as vibrato or legato playing, or identifying the gender and age of a speaker, are at the basis of describing the molecular elements of a stream of sound), or *Auditory Scene Analysis and Music Analysis* (interconnecting the individualized sonic elements of complex sonic environments into a global and abstract picture, where roles, functions, and relationships are defined).

Our work on sound analysis inside the SMS (Spectral Modeling Synthesis) framework [8] or sinusoidal models in general [9], and the tools we have developed so far, make possible to manage several measures that can be considered as being useful descriptors for the content of sound. They will be presented in the next section although, as we intend this paper to be of broad interest, we will not present the extraction methods. The descriptors we use can be computed in different flavors, depending on specific needs. Thus, we can consider an instantaneous value for them, a variation value for expressing differences between pairs of contiguous or separated frames, an average value, for describing content at longer time scales, and, associated with that average, a variance value, expressing the variability of the values at such a longer time scale. As it has been shown in relevant literature [2, 10] combinations of them can be used as a basis for elaborating higher-level descriptions that approximate ordinary ways of referring to sounds in search, classification, and transformation tasks.

Complementarily, audio content descriptions need a way for structuring and handling those descriptors together with the additional information that could be derived from them; we will discuss in another section the characteristics of such a *descriptor scheme*, considering that sound descriptions not only describe high level features (i.e. search for male voices, or for guitar solos), but also macro-structural and micro-structural (search for timbres with similar variations in the partial amplitudes along time, for example). As we will see, the description of an audio file with the help of spectral models can be done at multiple levels: we can describe it frame-by-frame, with so-called instantaneous or low level descriptors, but also can be done at a higher scale, taking into consideration a temporal *segment* of a file or stream, or a spectral *region* along a segment. We believe that the efficient exploitation of such a kaleidoskopical representation of sound can yield descriptions of (mainly musical) sound content that will be usable in multimedia applications.

2. Spectral models for sound description

There are several sound analysis techniques that can be used to obtain content descriptions (wavelets, LPC, cochleograms…) one of them being the spectral analysis based on a sinusoidal plus residual decomposition, SMS [11]. In this type of analysis, we compute the short-time Fourier analysis (STFT) of the sound, from which we derive a time-continuous representation of the sound in the form of sinusoidal tracks that follow the harmonic (or inharmonic) structure of the sound. This sinusoidal component is subtracted from the original sound, obtaining a residual signal that can be modeled with different spectral approximation techniques. Contrasting with other analysis techniques, as for example wavelets or LPC, such kind of representation is very intuitive from the point of view of a final end user of an audio database or audio processing software, and can be made quite more effective by deriving other attributes or descriptors in a hierarchical way, in order to preserve the information available at the lowest levels. Thus, we start with basic, and sometimes not quite semantic descriptions (as can be the "amplitude of sinusoidal track number three"), but we can end with mid- and high- level descriptors closely related to the ways we use to talk about sound (as can be "a sharp attack with a long sustained vibrato").

Low-level descriptors of sound content

The descriptors that we use constitute a set for a simple parameterization that accounts for the microstructure of a sound. In this set there are very basic parameters like *instantaneous frequency, amplitude and phase of each partial* and the *instantaneous spectral characteristics of the residual signal*. But, starting from them, there are also other useful instantaneous attributes that give a higher level abstraction of the sound characteristics, and that will be listed below. These attributes are easily calculated at each analysis frame from the output of the basic SMS analysis. We should also acknowledge that other relevant descriptors could be used, as for example odd/even partials ratio [12], tristimulus [13], attack harmonic coherence [14], and some of them will probably be incorporated soon in our system. For the moment this is our list:

- **Amplitude of sinusoidal component:** sum of the amplitudes of all harmonics expressed in dB.
- **Amplitude of residual component:** energy of the residual component expressed in dB.

- **Spectral shape of the sinusoidal component:** envelope described by the amplitudes and frequencies of the harmonics, or its approximation.
- **Spectral shape of the residual component:** approximation of the magnitude spectrum of the residual sound.
- **Harmonic distortion:** measure of the degree of deviation from perfect harmonic partials.
- **Noisiness:** measure of the amount of non sinusoidal information present in the frame. It is computed by taking the ratio of residual amplitude versus total amplitude.
- **Spectral centroid:** the midpoint of the energy distribution of the magnitude spectrum of the current frame. It could be considered as the “balance point” of the spectrum.
- **Spectral tilt:** the slope of the linear regression of the data points used to represent the spectral shape of the sinusoidal part.

Besides the instantaneous values, it is also useful to have parameters that describe the time evolution of an attribute. We describe it with the differences between frames.

Another important step towards a musically useful parameterization is the segmentation of a sound into fragments that are homogeneous in terms of certain sound attributes. Then we can identify and extract segment attributes that will give a summary of its content, and may allow to classify the segment into semantic categories corresponding to sound events or sound objects. One of the most obvious and general segmentation process divides a melody into notes and silences and then each note into an attack, a steady state and a release region. Global attributes that can characterize attacks and releases refer to the average variation of each of the instantaneous attributes, such as average fundamental frequency variation, average amplitude variation, or spectral centroid trajectory [15]. In the steady state regions it is meaningful to extract the average and variance of each of the instantaneous attributes and calculate other global attributes such as time-varying rate and depth of *vibrato* and *tremolo*. For more details on these expressive elements the interested reader can consult another paper from our group [16]. In summary, when we segment sound, we are using low-level content descriptors that can also help to characterize (when appropriately combined and interpreted) meaningful segments of sound. Contrasting with instantaneous measures, the segment description uses statistical measures such as mean andvariances in order to get the “global picture” along its duration.

As a final issue, it should be noted that the descriptors used in our spectral modeling environment allow a big degree of overlapping regarding its description power. This kind of overlap has also been used in existing systems for audiovisual content analysis [2, 10]. We believe that a standard for content description as MPEG-7 should accommodate this kind of redundancy, leaving to the front-end or content providers’ proprietary software the effective use of the descriptor’s set.

Mid-level and high-level descriptors

Describing sound at mid-level means determining sound events and objects. Although the distinction between events and objects can be a little bit controversial, we will consider that a sound object is a sound source, and that any kind of behavior of that object is an event. Events develop in time, so all events have a duration property. Although in a Schaefferian sense [17] all events can be considered objects because they can be grasped through an operation of “reduced hearing”, from a functional point of view it seems more convenient to separate sources from their behaviors, as we do.

Moreover, describing sound at what we consider “high level”, means incorporating events and objects into formal structures that convey musical meanings and roles for them. Examples of such a kind of structures can be found in the implication-realization by Narmour [18], or in the generative theory by Lehrdal & Jackendoff [19], but other non-musicological structures can be also considered as high-level. As it is a matter clearly outside the current scope of automatic procedures, and much a cognitive/musicological issue than an engineering one, we will keep it aside. On the other hand we will concentrate in mid-level descriptions, as it is still currently an active area of basic and applied research.

Regarding the mid-level descriptions of sound, the most obvious description is –at what we could call alongside Rosch [20] the “basic level” of categorization- the description into notes (in case of speech it could be the description into words). At a lower level we need to decompose notes into envelope stages as attack/steady state/release, and at a higher one we need to group notes into phrases and melodies. Melody identification and representation [21] is one of the most important problems faced by content-based multimedia databases and it has given birth to an interesting new search modality called “query by humming” [22, 23]. Recent incorporation of rhythmic constraints will improve the performance of that kind of systems [24].

Identifying sound sources is another “basic level” categorization task, but a very different one in nature. Although we can use the same set of descriptors than for identifying events, describing sound sources is equivalent to describing a perceptual property: timbre. As timbre is a complex perceptual dimension (i.e. not as unidimensional as loudness or pitch), there are several variables and dimensions that determine one timbral sensation or another [25]. We can quickly grasp the complexity of this problem by noticing that two sounds generated by the same sound source, but separated two octaves in pitch do not share the same attributes that were shared when notes were very close. Thus, automatic processes for identifying sound sources are not quite effective yet, although there are scattered interesting results [26] [27]. For these reasons, similarity based searches and indexings of sounds are still quite diverse and controversial. Anyway, systems as SoundFish [28] or the Studio On Line developed at the IRCAM [29] show that effective features and procedures do exist for quite well solving those tasks.

Once a source has been identified we can ask for attributes like materials or maker (for example, once we have identified a sound as one of a piano the description should be completed with attributes like if it is a vertical or a grand, a Steinway or a Yamaha, etc.). We can also ask for more general attributes such as the instrumental family. As recent work reveals [30], it could be easier to start with this level of identification instead of trying to directly identify the specific sound source. Another kind of description related to the source is the acoustic environment where sound is produced, that can also described in a systematic way [31].

3. A Sound Description Scheme

As a way of organizing not only our own descriptors but also others that might be proposed, we have devised what in MPEG-7 jargon is called a “Description Scheme”, that is, a structure that specifies the semantics and the relationships between Descriptors and other Description Schemes. The scheme started as an extension of SDIF (Sound Description Interchange Format) [32] [33], an ongoing proposal devised through the collaboration of IRCAM, CNMAT and IUA sound laboratories, that was intended for the storage of spectral analysis data. Since then it has evolved into a thorough scheme inspired not only in our knowledge about sound but also on proposals made for visual schemes in the context MPEG-7.

The scheme we present here addresses the description of multichannel, multi-source sounds. It is intended to be a generic description scheme based on a hierarchical (and recursive in some places) structure, that can describe sound at multiple levels of detail (from the low level of an FFT analysis frame to the high level of a whole sound file), accommodating different kinds of dependencies and relationships among its components. Given its modularity and expandability it can be used to describe sound in general (although speech description will require more specific schemes), and it can take into account different kinds of descriptors of sound. The scope of descriptions can be changed as needed and depending on the target application, and there are descriptors and description sub-schemes that are not compulsory at all. This way, the scheme can contain descriptions at only one level, or at all available levels of description.

The scheme has two main levels: *syntactic* and *semantic*. We need a syntactic level in order to describe the sound file or stream in a superficial, temporal-structural way. But we also need a semantic level in order to assign semantic labels to the elements that deserve attention. The syntactic level is used to specify physical structures and the signal properties of the sound program. The elements of the syntactic part of the scheme are *tracks*, *segments*, specific instances of segments like *frames*, and *regions*. On the other side, the semantic level is used to specify semantic features of a sound program in terms of *audio events* and *audio objects*.

Other elements at its topmost level are: a Model DS for describing the analysis and classification data in a compact and abstract way, a Syntactic-Semantic Link DS for describing relationships between elements of both levels, a Media DS for describing respectively storage format and other technical characteristics, Meta DS for manual descriptions, and a Summary DS for quick audio and visual browsing of relevant information.

Description Schemes for the syntactic level

Track Description Scheme

A first substantial difference between Video media and Sound media is that sound can be expressed along several channels at the same time (stereo recording, Dolby Surround, multi-track recordings…). Those channels may contain related information but also very distinct ones so it is clear that we need a scheme to describe each track separately (while allowing links between track descriptions).

A track has no more the length of the whole audio program, although it could be shorter. Tracks do not need to start at the same time, although it could be usual to do so. In case the program is mono only one track is needed. In case of being multichannel, one track is provided for every audio channel. Certain situations where only one track could be used with multichannel audio could be envisioned (when double mono channels are stored in one stereo file, or when content in two or more channels is supposed to be practically identical...).

A track DS is then described by:

- a **Time DS**, that describes the beginning and the end of the track (time relative to an absolute time of the whole file).
- a **Placement DS**, which describes the spatial placement of the track from the listener point of listening (i.e. left, right, 60 degrees left, behind, etc.) for the track.
- a **Segment DS**, which contains the different segments created by applying different segmentation criteria to the track, alongside additional segmentation information
- a **Region DS**, that describes the microstructure of the track
- a **Track Linking DS**, that describes if the track is linked to another track (i.e. shares the same content descriptions) or any other kind of interaction or relationship between tracks. When tracks are linked they share the same Segment DS (for example stereo-recordings where both channels are related do not need separate Segment DSs. In that case we would like to benefit from the description of the left Track for the right Track)
- a **Summary DS**, that describes different ways for quick listening or "auralization" of contents, and also for fast browsing of audio contents (as musical score, midi file, spectrogram…)
- a **Media information DS**, which describes information specific to the storage media (i.e. sampling rate, resolution, format, compression format, etc.)
- a **Meta information DS**, which contains information that usually cannot be extracted from the signal itself (title, author, technical crew, date…)

Note that the last three DS's do not appear in Figure 1 at the track level for keeping the graph more clear.

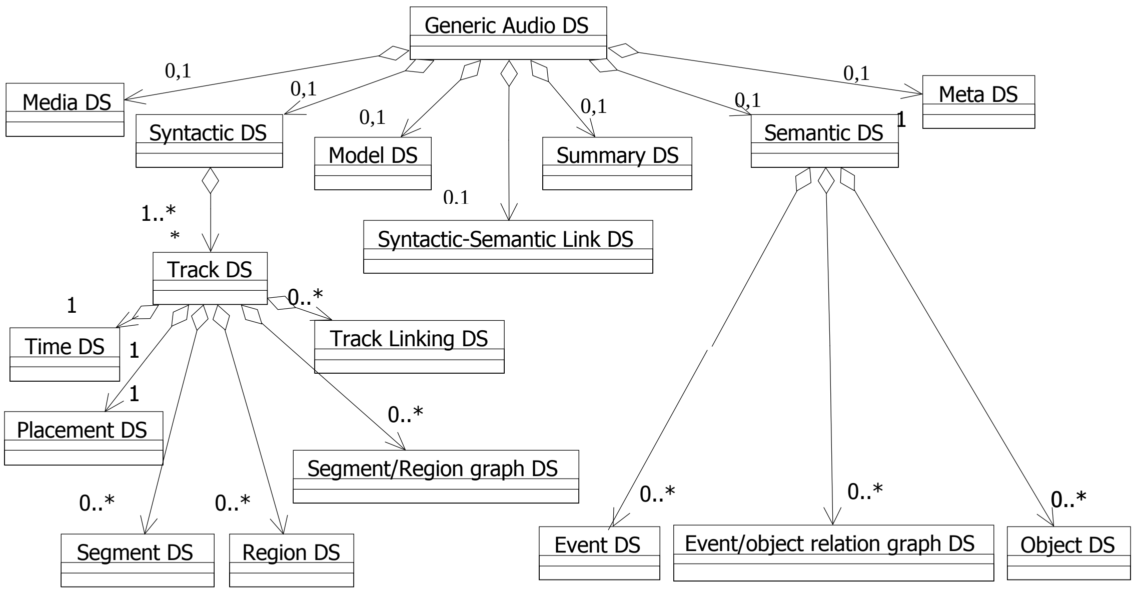

Fig. 1. Overview of Audio Description Scheme

Segment Description Scheme

An audio program can be segmented into one or more groups of contiguous samples that share some common property (for example, utterances of the main character of a video film, the chorus of a song, a vibrato note of a soprano...). Each one of such groups is an audio segment. Audio segments can be further segmented into another group of segments, yielding a tree segment. A track can contain different segment trees in order to accommodate different overlapping segmentations according to different criteria (i.e. by speaker gender and by background music/not background music). General hierarchy of the

Segment Tree goes from a root segment into subsegments, sub-subsegments, and finally into frames, which consist of samples.

In summary, the Segment DS is described by:

- a **Time DS**, that describes the beginning and the end of the segment
- a **Media information DS**, a **Meta information DS**, an **Auralization DS**, and a **Visualization DS**, as defined in the track DS
- a **Segmentation Criterion**, with the name for the criterion used in the segmentation
- a **Segment-Region Link DS**, allowing to link the segment with region descriptions (spectral low-level features)
- several optional **Temporal Descriptors**, that describe temporal low-level features of the sound (i.e. autocorrelation, zero-crossing rate, etc.)

Lowest levels of a segment tree can be:
the **Frame level**: a frame is usually the shorter meaningful Segment DS. It allows a deep description of the local sound characteristics through spectral analysis (or other different technique). A frame has exactly the same properties than other Segment DSs. As it is the lowest level of the Segment DS the descriptors attached to this level represent instantaneous values.
the **Sample level**: samples are like atoms of the sound, and except in special cases like glitches, they do not carry any meaning or content by themselves. Neither further decomposition of them is possible nor linking them to a Region DS. The reason for including this level is that sometimes can be useful to have a sample-by-sample segment description.

Region Description Scheme
It describes the microstructure of a segment of sound, in the form of a sound spectrum, or a series of sub-spectra inside the effective bandwidth of a sound. The region can be linked with any kind of segment (a long one or a frame-level one). That opens the possibility for a description of sound ranging from the microstructural level (when the Region DS is linked to a frame-level Segment) to the macrostructural level (when the Region DS is linked to a higher-level Segment). There can also be several region decompositions associated to the same segment, and it is also possible that several segments share the same region decomposition.
A Region DS is described by:

- a **Spectrum DS** , which describes the main spectral features of the sound
- an optional **Motion/Deformation DS**, which describes the evolution of the Audio Spectrum along a short period of time

Spectrum Description Scheme
A Spectrum DS consists of:

- a series of **Specific Global Descriptors** like *amplitude*, *fundamental frequency*, *spectral range*, *noisiness*, etc. The value for this descriptors is
  - an **instantaneous** one when the Region DS to which it is connected is linked to the frame level of the Segment Tree
  - a **difference** one when the Region DS to which it is connected is linked via a Region Motion/Deformation DS to the frame level of the Segment Tree
  - a **mean** one when the Region DS to which it is connected is linked to a higher level of the Segment Tree
  - a **variance** one when the Region DS to which it is connected is linked via a Region Motion/Deformation DS to a higher level of the Segment Tree
- a **Spectral Shape DS**, that describes the energy profile across the spectrum of the frame both in a global way with a *Spectral Envelope Profile DS* (containing an envelope, LPC coefficients, Mel-cepstrum coefficients, formants, etc) but also with **optional descriptors** used to describe specific features of the spectral shape (most of them were listed in the low-level descriptors section, as for example: Spectral Centroid, Spectral Tilt, Noise Shape Harmonic Distortion, Odd/Even ratio, etc.).

Description Schemes for the Semantic level

In the semantic part we describe two broad categories of sonic elements: events and objects. Given the temporal nature of audio events, they are mainly linked with segments described in the syntactic part. Likewise audio objects are closely (but not exclusively) linked to the regions described there. Events can be considered as temporal constrained behaviors of objects. Objects, then, are the generators of the events.

Event Description Scheme
An event is the temporal behavior of some audio object along or around a certain segment of time. Typical audio events can be: a melody, a musical phrase, a "solo" section, a musical note, different sections of a note regarding its amplitude evolution (attack/steady state/release), an audio fade (in or out), a sentence uttered by somebody in a video, the words in the previous sentence, the phonemes that the previous words are made of, non-linguistic utterances (crying, shouting, sighing, etc.)... An Event DS can contain an arbitrary number of Event DS. Therefore Event DS form a Tree (for example: a musical motive or melody is composed by different phrases, these phrases are composed by different notes, and the notes have different envelope sections). Of course, there can be more than one tree. Therefore, in the Event Tree, we won't talk about "*violin*" or "*cello*" (which are objects) but about their temporal behavior such as "*violin phrases*", or "*cello notes*".

An Event DS is described by:

- an optional **Annotation DS**, which is a text descriptor for describing non-automatically-extractable features
- an **Event Division Criterion DS**, which describes the criterion used to divide the Event in Sub-Events
- an **Event-Segment Link DS,** which links one Event with one or several Segments
- a set of **Temporal Descriptors**

Object Description Scheme
The main function of an Object DS is describing sound sources. It is possible to distinguish different levels for describing audio objects. The most generic source objects can be: musical instrument, speech voice and environmental sound/sound effects. For musical instruments more detailed levels can be: musical family of the instrument, specific instrument, excitation resonance characteristics (type and structure of excitation, resonance structure), material that is made of, specific shape characteristics, manufacturer and model, acoustic environment where they sound, etc. For speech voice: gender, age segment (child, young, mature, old), excitation-resonance characteristics, identity of speaker, acoustic environment where it sounds, etc. Finally, for environmental sounds: space/time trajectory of the source, excitation resonance characteristics, specific source, etc.

An Object DS contains an arbitrary number of Object DS and therefore form a tree (i.e. a musical ensemble can be an object made of groups of instruments –strings, reeds, brass, voices, soloists-, and these groups can be, in turn, made of small subgroups –strings: violins, cellos, violas; voices: sopranos, tenors, etc.-, down to the individual instrument, if necessary).

An Object DS is described by:

- An optional **Annotation DS** with text descriptions extracted by hand
- An **Object Division DS**, which describes the criteria used to construct the object tree
- One or more **Object Type Descriptors**
- One or more **Object Behavior DS** which describe the object behaviors (they link to events)
- **Object Interaction DS** which describes interactions between different objects (i.e. the bow and the body of a cello, a singer's voice and her microphone…)
- an optional **Object-Region Link DS**, for linking the object with its relevant spectral information
- an optional **Object-Event Link DS**.
- **Object Descriptors** which describe the sound sources at different levels as discussed above.

4. Conclusions

Spectral models are suitable for describing sound at different levels of abstraction. Low level descriptions can be used as building blocks for mid-level and high-level descriptions, or for models [34] that allow to derive those descriptions. Starting from spectral models representational framework we have developed a description scheme for audio in MPEG-7 that mainly encompasses low-level and mid-level audio descriptors, but can also accommodate high-level descriptors (although they have been kept outside the scope of this discussion because they should be handled in a non automatic way). From a user point of view, MPEG-7 is a challenging initiative that should improve our efficiency for accessing multimedia contents. But it is also challenging from an academic and engineering point of view because it addresses problems that are still hot research topics for the audio community, and therefore they must be solved in order to provide a useful and long-life standard for multimedia content description.

References

[1] MPEG 7: MPEG-7: Context, Objectives and Technical Roadmap, ISO/IEC JTC1/SC29/WG11/N2729 V 11. Seoul meeting, March (1999)

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

[3] L. Wyse, Smoliar, S.W.: Toward content-based audio indexing and retrieval and a new speaker discrimination technique. In: Rosenthal, D. F., Okuno, H. G. (eds.) Computational Auditory Scene Analysis: Proceedings of the IJCAI-95 Workshop. Erlbaum (1998)

[4] Martin, K. D.: Toward automatic sound source recognition: identifying musical instruments. In: Proc. NATO Computational Hearing Advanced Study Institute, Il Ciocco, Italy, July 1-12 (1998)

[5] Foote, J. T.: A similarity measure for audio classification. In: Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora, Stanford, CA, March (1997)

[6] Zhang, T. and Kuo, C.J.: Content-based Classification and Retrieval of Audio. In: SPIE's 43rd Annual Meeting – Conf. on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII, SPIE Vol.3461, San Diego, July (1998) 432-443,

[7] Scheirer, E.D.: Towards music understanding without separation: segmenting music with correlogram comodulation. In: Proc. IEEE Workshop on Signal Processing to Audio and Acoustics, Mohonk, NY. (1999)

[8] Serra, X., Bonada, J.: Sound Transformations on the SMS High Level Attributes. In: Proc. DAFX98: First Digital Audio Effects Workshop. Barcelona (1998)

[9] Peeters, G., Rodet, X.: Signal Characterization in terms of Sinusoidal and Non-Sinusoidal Components. In: Proc. DAFX98: First Digital Audio Effects Workshop. Barcelona (1998)

[10] Pfeiffer, S., Fischer, S., Effelsberg, W.: Automatic audio content analysis. In: Proc. ACM Multimedia 96. Boston, MA. November (1996) 21-30

[11] Serra, X.: A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. PhD dissertation. CCRMA, Stanford University, CA (1989)

[12] Kostek, B., Wieczorkowska, A.: Parametric representation of musical sounds. Archives of Acoustics. 22(1*)*. 3-26. (1997)

[13] Pollard, H.F, Janson, E.V.: A tristimulus method for the specification of musical timbre. Acustica, 51 (1982)

[14] Grey, J. M.: Multidimensional perceptual scaling of musical timbres. J. of the Acoust. Soc. of America, 61, (1977) 1270-1277

[15] Hadja, J., Kendall, R., Carterette, E., Harshberger, M..: Methodological issues in timbre research. In: Deliège, I., Sloboda, J. (eds.) Perception and Cognition of Music. East Essex: Psychology Press (1997)

[16] Herrera, P. and Bonada, J.: Vibrato extraction and parametrization in the Spectral Modeling Synthesis framework. In: Proc. DAFX98: First Digital Audio Effects Workshop. Barcelona (1998)

[17] Schaeffer, P.: Traité des objets musicaux. 2ème édition. Paris. Éditions de Seouil (1977)

[18] Narmour, E.: The Analysis and Cognition of Melodic Complexity: The Implication-Realization model. Chicago, University of Chicago Press (1992)

[19] Lehrdahl, F., Jackendoff, R.: A generative theory of tonal music. MIT, Cambridge, MA (1983)

[20] Rosch, E.: Principles of Categorization. In: E. Rosch and B.B. Lloyd (Eds.), Cognition and Categorization. Hillsdale, NJ: Lawrence Erlbaum (1978)

[21] Lindsay, A.T.: Using contour as a mid-level representation of melody. MSc thesis. MIT, Cambridge, MA (1996)

[22] Ghias, A., Logan, J., Chamberlin, D., Smith, B.C.: Query by humming: musical information retrieval in an audio database. In: ACM Multimedia 95- Electronic proceedings. San Francisco, CA. November (1995)

[23] McNab, R.J., Smith, L.A., Witten, I.H., Henderson, C.L., Cunningham, S.J.: Towards the digital music library: tune retrieval from acoustic input. In: Proc. ACM Digital Libraries 96. March (1996)

[24] Handel, S.: Timbre perception and auditory object identification. In: Moore, B.C.J. (ed.). Hearing. San Diego, CA: Academic Press (1995)

[25] Shmulevich, I., Yli-Harja, O., Coyle, E., Povel, D., Lemström, K.: Perceptual issues in music pattern recognition-Complexity of rhythm and key finding. In: Proc. of Symposium of Artificial Intelligence and Musical Creativity. Edinburgh (1999)

[26] Brown, J. C.: Computer identification of musical instruments using a pattern recognition with cepstral coefficients as features. Journal of the Audio Engineering Society, 105(3*)* (1999) 1933-194

[27] Martin, K.D.: Sound-source recognition: a theory and computational model. Ph.D. Thesis, MIT. (1999)

[28] Blum, T., D. Keislar, J. Wheaton, and E. Wold.: Audio databases with content-based retrieval. In: Maybury, M.T. (ed.) Intelligent Multimedia Information Retrieval. Menlo Park, CA: AAAI/MIT Press (1997)

[29] Ircam-SOL Search Engine page: http://www.ircam.fr/equipes/instruments/sol_psy/Scripts_english/cuidad4.html

[30] Martin, K.D. and Kim, Y.E.: Musical instrument identification: A pattern-recognition approach. In: Proc. of the Acoustical Society of America. October (1998)

[31] Jullien, J.P., Kahle, E., Winsberg, S., Warusfel, O.: Some results on the objective and perceptual characterization of room acoustical quality in both laboratory and real environments. In: Proc. Institute of Acoustics, XIV (1992)

[32] Wright,M, A. Chaudary, Freed, A., Wessel, D., Rodet, X., Virolle, D., Woehrmann, R., Serra, X.: New applications of the Sound Description Interchange Format. In: Proc. ICMC-98, Ann Arbor, MI (1998)

[33] Ircam-SDIF Web page: *http://www.ircam.fr/produits/techno/multimedia/Cuidad/SDIF-e.html*

[34] Casey, M.A.: Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio. Ph.D. Thesis. MIT Media Lab (1998)