Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, July 2-5, 2002

SCALABLE METADATA FOR SEARCH, SONIFICATION AND DISPLAY

Alain de Cheveigné

Ircam - CNRS
1 place Igor Stravinsky, 75004, Paris
cheveign@ircam.fr

ABSTRACT

This paper argues for the need - and usefulness - of scalable content-based metadata. Scalability is here defined as the conjunction of two properties: arbitrary resolution, and convertibility between resolutions. The need follows directly from the projected exponential trend of media data size, that equally affects metadata. In addition to addressing this need, scalable metadata are useful because they are hierarchical in nature, and incorporate statistics effective for search (in automatic media handling systems) or sonification and display (in interactive media handling systems). Scalable metadata are built upon a small number of statistical operations that offer the right scalability properties: extrema (min, max), mean, variance, covariance, histogram, etc. These statistics are used alone or in combination to produce summary descriptions with a resolution tailored to the needs and constraints of the application. They can also be understood as parametrizations of the distributions of full-resolution descriptor values that they summarize. As such, they support inference mechanisms upon to build search and matching algorithms. For interactive applications, scalable content-based descriptors can be used to produce visual displays that support zooming and navigation within multimedia collections of arbitrary size, under the assistance of visual and auditory feedback.

1. INTRODUCTION

This paper centers on the concept of scalable metadata. Scalability (defined below) is a property that metadata must have in order to fill their role over a significant time span. Scalability aims to avoid the development of successive layers of "meta-" metadata in response to the exponential increase of metadata size. Scalability puts particular constraints on the design and semantics of descriptors. In return, scalable metadata have useful properties that can help to build indexing, search and comparison algorithms, as well as hierarchical structures for navigation.

In the context of this paper, "metadata" are understood to be numerical and derived automatically from content. Examples for audio might be fundamental frequency (F_0) or spectrum. Similar issues exist for other metadata such as hand-input text annotations, but scalable operations are harder to define for these. Arguably, after sufficient summarization, most metadata must become numerical (e.g. the symbolic tag "artist name" leaves place to the *number* of documents by that artist).

1.1. The need: handling the growth of metadata

The exponential growth of data is a well known phenomenon, sometimes referred to as "Moore's law" by analogy to a similar trend for the density of integrated circuits observed by Gordon Moore in 1965 and verified since then [2]. Magnetic storage density and sales volume tend to double every year [3], as does the bandwidth of networks [1, 4]. Estimates of "world wide web size" are more approximate, but it is likely that the rate of growth is as large or larger. This trend affects all sorts of data including multimedia data.

As a consequence, consumers, creators and administrators confront ever-larger collections of ever-larger (or ever-richer) documents. Examples are more TV channels, larger catalogues, new services (such as access from mobiles), or new "high resolution" or "interactive" media formats. At the same time, the cognitive and behavioral bandwidth available to consume or manipulate the content remains more or less constant. The result is an ever-widening gap between the scale of the user and that of the data[1].

This problem is well known, and it is a driving force behind the concept of metadata ("the bits about the bits"), and initiatives such as MPEG-7 to promote the emergence of effective standards for metadata [5]. Metadata serve to "represent" the data for operations that do not require access to the content. Supposing that they are more compact or better organized than the data, they are expected to facilitate manipulations by the user (or by the software agents used by the user).

Metadata are a step towards a solution, but they carry the seeds of a new problem. If metadata are attached to all data, and data increase exponentially, so must metadata. Metadata appropriate for small documents may be too cumbersome for large, while metadata for large documents may be inappropriate for collections, etc. With each new tool comes new formats, abstractions, and user interfaces. This poses two sorts of problem: technical and human. The first stems from the accumulation of generations of "legacy" metadata, with their cost in bulk and interoperability problems. The second results from the succession of new interfaces and abstractions that a user must learn with each new layer of metadata. This is all the more annoying as the task of organizing content, by its nature, calls for seamless interoperability.

1.2. Scalable metadata

The concept of scalable metadata was developed to address these issues. By "scalable" we mean two properties. First, the metadata format must allow a description to be instantiated at any resolution. Second, an existing description must be convertible automatically to a lower-resolution format, with the guarantee that the result depends only on the resolution and not on intervening scaling operations. Of course, each scaling operation causes a *loss of informa-*

[1] An interesting correlary is that much content must consist of copies. "Creative bandwidth" does not increase that fast.

tion. This is the price to pay for conciseness, imposed by storage, bandwidth or cognitive constraints. We require convertibility from high to low resolution, obviously not the opposite.

Scalability requires well-defined operations to transform one resolution to the next. It turns out that several mathematical operations have the right properties. Among them are extrema (*min* and *max*), *sum*, *mean*, *variance* and *covariance* (if stored together with the mean), and *histogram*. Scalable numerical metadata consist of descriptor values summarized by one or several of these operations, together with various bookkeeping information.

Scalability is best taken into account in the design of descriptors themselves. One reason is that scaling operations are more easily implemented at the systems level, operating on elementary "scalable" data structures, than at the application level operating on complex descriptors. Another is that, if the extraction algorithm is homogenous with a scaling operation (e.g. audio power and power spectrum are homogenous with the mean operation), then the descriptor semantics will be uniform across scale [29, 30, 28].

1.3. Interoperability and the life-cycle of metadata

Metadata are a very different beast from content. Taking metaphors from biology, they are expected to diffuse more easily, have a longer life cycle, and be more persistent. For example a vendor might distribute content for a fee, but metadata for free, or software agents might collect metadata systematically, but data only on demand. Content may be transcoded and the old format discarded, but vintage metadata may survive longer, because they are both more complex and less costly to keep.

Metadata address a task that is encyclopedic in nature: they must handle documents, collections, collections of collections, etc., including data from various sources. A good metadata-consuming agent should accept *every* format. Most metadata are likely to be produced by one application and consumed by another. Metadata are thus destined to be reused, often for purposes different than that for which they were created. One can predict an opportunistic behavior for metadata-consuming agents, and a long life-cycle for metadata. For these reasons *reusability* and *interoperability* are important, and this justifies standardization initiatives such as MPEG-7 [5].

An important aspect of the life-cycle involves changes in *scale*. Metadata may start their life produced by an application that requires and can afford relatively high resolution (for example a production editor). As time goes by, requirements and constraints may change so that it no longer makes sense to maintain full resolution. Deleting is drastic: it is much better to allow the metadata to be gracefully "shrinked", preferably automatically under system control. Alternatively, the problem may be simply that the metadata store is too vast for a human to browse, the solution being to percolate up enough information to derive a synthetic view that is easier to grasp. Or the user may be operating via a slow network. Scalable metadata address such needs.

1.4. Search, sonification and display

It is fruitful to distinguish between automatic and manual operations. A major use for metadata is as an index to expedite operations such as content-based search, classification, stream monitoring, duplicate detection, etc.. These could in principle be done directly on the content, but only if it is online, and the costs may be high. Efficient search pivots on pruning or prioritizing the search to reduce the time taken to terminate. Scalable metadata fit well within tree-based search structures, and offer parametrizations of the distributions of descriptor values summarized at each node. The presence (or better: absence) of a target value within a subtree can be known without actually visiting it, from the statistics attached to its root [25, 26, 8, 10].

Audio editors typically use a combination of spectrogram and waveform displays. For all but the smallest documents, there are usually more spectrum frames than pixels horizontally, or frequency bins than pixels vertically. Plotting them all is wasteful and unnecessary. Furthermore, extracting features to be displayed directly from content may be time-consuming or impossible if the data are off-line, whereas constraints on storage or handling time may preclude the storage of precalculated full-resolution features. Scalable metadata are of use here, as resolution may be tailored neatly to the various (and time-varying) constraints of display, transfer and storage.

Audio features such as spectrogram, F_0, harmonicity and modulation spectrum support visual display, but they may also be used to synthesize summary "earcons" to allow quick auditory browsing. The same metadata may be shared among search, sonificiation and display, and even the stuff of a desktop icon may be of use for all three purposes.

The following section describes a set of scalable summarization operations. Subsequent sections describe their application to audio descriptors, and give examples of search, display and sonification. Finally some comments are made on scalability within MPEG-7.

2. SUMMARIZATION OPERATIONS AND SCALABLE DATA STRUCTURES

Consider a full-resolution description consisting of N descriptor values. Supposing that this set is partitioned into subsets of K_i values each, we consider operations that map each subset to a single value.

An elementary summarization operation is *sum*, defined simply as:

$$s = \sum_{j=1}^{K} x_j$$

where x_j are samples to be summarized. It is obvious that K can be arbitrarily large and thus descriptions arbitrarily concise. Associativity insures that the result is the same if the operation is applied first to each subset of a partition of the original subset, and the results then summed. Descriptions scaled by sum are therefore scalable as defined in the Introduction. A useful variant is the *weighted sum*:

$$s = \sum_{j=1}^{K} w_j x_j$$

where w_j represents the *weight* of x_j, which may be used to modulate its contribution, for example according to reliability. Weights themselves are scaled by the sum operation:

$$w = \sum_{j=1}^{K} w_j$$

To be scalable, a description must include both values (or sums) and weights (or summed weights).

The weighted *mean* is likewise defined as:

$$m = \sum_{j=1}^{K} w_j x_j / \sum_{j=1}^{K} w_j$$

If each sample x_j has a weight of 1, the definition defaults to that of a unweighted mean (the weight of m is then K). To be scalable, a description must include both means and weights.

Operations *min* and *max* also allow scalability. Optionally they too may involve weights, with the convention that zero-weight samples are ignored when calculating the min and max. *Deterministic decimation* (take one in K) likewise ensures scalability (weights are applied as for min and max). *Probabilistic decimation* (take one in K at random) supports a similar but weaker property (identity of statistical properties, rather than of values, between same-scale descriptions of the same data). Weights (or counts) determine probabilities at each decimation. The *histogram* operation (*quantize* and *count*) produces descriptions that can be scaled by applying the operation *sum* to bin counts. If weights are present they are used in place of counts. The quantization rule must be attached to the description.

Weighted *variance* (for vectors: *covariance*) yields a scalable description if associated with the mean and weight. Weighted variance may be defined as:

$$v = \sum_{j=1}^{K} w_j (x_j - m)^2 / \sum_{j=1}^{K} w_j$$

where m is the weighted mean. When rescaling, the variance v of an aggregate of subsets can be derived from their variances, means and weights (v_j, m_j, w_j) according to the formula:

$$v = \sum_{j=1}^{K} w_j v_j / \sum_{j=1}^{K} w_j + \sum_{j=1}^{K} w_j (m_j - m)^2 / \sum_{j=1}^{K} w_j \qquad (1)$$

For vectors, variance is replaced by covariance matrices. To save space these may be summarized by their diagonal or trace, each of which yields a scalable operation. Before scaling it may be useful to apply a linear transform (for example based on PCA) to diagonalize the covariance matrices, so that the series of matrices can be accurately represented by a series of diagonals. The transformation matrix is then attached to the description.

The operations just described can be interpreted as *statistics* on the population of values that they summarize. Each provides useful information reflecting some aspect of that population. Deterministic and random decimation retain actual exemplars. Extrema give deterministic bounds on the population extent. The histogram measures the density within each quantization category. The mean measures central tendency, and the variance and covariance (together with the mean) measure extent. They can be used for example to parametrize multivariate gaussian models of the population (trace of the covariance matrix: hyperspherical, diagonal: ellipsoidal aligned on axes, full matrix: arbitrarily oriented ellipsoidal).

A more precise description of the "granularity" of the distribution can be obtained by decomposing each variance coefficient into a sum of terms, each of which reflects variance at a certain scale. This is done by first calculating the total variance. Samples are then averaged two-by-two, and the variance calculated again. The difference between the two yields a first term that reflects variance at the finest scale. Averaged samples are then averaged again, and a third variance estimate is derived and subtracted from the second to obtain a second term reflecting variance at the next-to-finest scale. A total of $q = log2(K)$ terms may thus be obtained. Their sum equals the variance of the distribution. This decomposition, dubbed *scalewise variance* is scalable. If the descriptor values form a time series, their scalewise variance has properties similar to an *octave-band spectrum*.

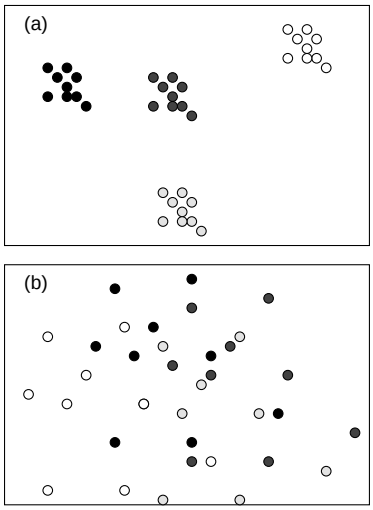

Fig. 1. The two distributions have equal total variance but different scalewise variance vectors. In (a) the samples of each subset (distinguished by darkness) are tightly distributed but the subsets are widely distributed. Low-order coefficients of scalewise variance are small, high-order coefficients are large. In (b) the samples within each subset are widely distributed and the subsets. Low-order coefficients of scalewise variance are large, high-order coefficients are small.

Content-based descriptors of audio or video media typically occur as time-series. If scaling operations are constrained to use subsets that are contiguous, the temporal structure is conserved by scaling. The scale ratio K may be uniform across the scaled series, or it may vary to provide non-uniform resolution. Once scaled, the values are stored in a *data structure* that holds the series of scaled values, eventually their weights, the count of samples that each scaled sample reflects, and a tag specifying the scaling operation that was applied. See the section on MPEG-7 for details of an actual implementation.

In summary, a number of operations are available to build scalable descriptions. It is possible to use several operations for the same descriptor (e.g. min and max, mean and variance), possibly with different temporal resolutions (e.g. a high resolution series of means combined with a low-resolution series of covariance matrices). This variety allows catering to the needs of a range of applications. It also raises interoperability issues that are discussed in the section on MPEG-7.

3. SCALABLE AUDIO DESCRIPTORS

This section shows a few examples of scalable audio descriptors, some of which have been included in the audio part of the MPEG-7 standard [28, 19].

3.1. Waveform

The simplest "descriptor" of audio content is the content itself (waveform), although it seems strange to call it "metadata". It makes more sense once the data are scaled by appropriate summarization operations. The min and max operations produce waveform descriptions useful for display and search (see below). The histogram operation may also be useful, whereas operations such as decimation, mean or variance are less appropriate (they certainly don't display well). Note that the aim is *not* to allow perceptually acceptable restitution, as with coding schemes.

3.2. Power

Power is the mean over an interval of the squared waveform. This makes it homogenous with the mean summarization operation: descriptions scaled by mean keep a uniform semantics across the range of scales. It may be useful to also include variance, and in particular *scalewise variance*, which decomposes the variance over several scales, and thus offers information analogous to a *modulation spectrum*.

3.3. Fundamental frequency

F_0 is a good predictor of the perceptual dimension of pitch. It is useful to interface between content and high-level melody descriptions, for search and classification (in combination with spectrum), and for display. A problem with F_0 is that it is not always defined, and thus the F_0 time series is peppered with "garbage" values. If no steps are taken to discount them in scaling operations, they corrupt the scaled samples. This is where *weight* is of use: each F_0 sample is shadowed by a weight sample (quantized to 0-1 or graded) that determines its contribution to scaling operations. F_0 time series can thus be effectively scaled. The *histogram* operation is useful to describe the density of occurrence of each pitch (or chroma) class.

3.4. Spectrum

A spectrum descriptor is also useful, the main difficulty being choosing among the many ways of calculating it. Power spectra that follow a logarithmic, Bark or ERB frequency scale are a good choice. These scales tend to distribute power evenly over channels, and their frequency resolution approaches that of the ear. Power is homogenous with the mean operation, which gives the descriptor uniform semantics across resolutions when that operation is used for summarization. Other representations may also be useful: log power, cubic root, etc., or transformations such as cepstrum, PCA, ICA, etc.. Uniform semantics are lost in that case, but the scaled representation (e.g. mean plus variance or covariance) remains useful to parametrize models of the distribution of spectrum values.

3.5. Other descriptors

Other useful descriptors are *harmonicity* (defined based on the amount of inharmonic power, and represented either directly or as a ratio to total power), *modulation spectrum* (the spectrum of the time series of instantaneous power of the signal or individual frequency bands), spectral centroid, spread, kurtosis, etc.. These too can be made scalable. All of these descriptors describe continuous features, but the case may be made for a different class of features based on "events" (onsets, silence, etc.) for which the histogram provides a scalable representation. To summarize, major content-based audio descriptors can readily be made scalable.

4. SEARCH AND COMPARISON

Efficient search means *pruning*. This requires inferring (on deterministic or probabilistic grounds) that a subset of the search space does not (or is unlikely to) contain the target [25, 26, 8, 10, 8]. The sooner such a decision can be made, the quicker the search terminates. Scalable metadata help in two ways: by supporting a hierarchical metadata structure that allows pruning high in the hierarchy, and by offering at each node statistics that describe the subset that they summarize. Statistics such as min, max or histogram allow deterministic pruning [9]. Mean, variance and covariance can be used to parametrize distributions (such as gaussian) to support probabilistic pruning [26]. Scalewise variance allows search decisions to be made on the basis of the granularity of the underlying population (a "lumpy" population is cheap to search, a diffuse population expensive).

It is not the aim of this paper to review the many search methods that can make use of scalable statistics. Two recent examples are the multiple speaker detection method of [13] based on the Bayesian Information Criterion that uses a formula analogous to Eq. 1 to hierarchically aggregate covariance matrices, or the efficient audio search method of [14] based on histograms of occurrences of vector-quantized spectra.

As an simple illustration, consider how extrema statistics can be used to quickly compare audio waveforms (for example to search for duplicates on a hard disk, or find the file from which was extracted a clip). Every file is supposed to be labeled by a series of min, max pairs down to some resolution. To compare two files, the algorithm first builds for each file a hierarchical tree of extrema over progressively larger intervals (this tree may also be precalculated and stored within the metadata). It then compares intervals of both files, starting at the coarsest level. Consider intervals A and B from either file, with minima m_A and m_B and maxima M_A and M_B. The following inferences can be made: if $m_A > M_B$ or $m_B > M_A$ the intervals are distinct (neither contains a subset of the other). If $M_A > M_B$ or $m_A < m_B$, then A is not a subset of B. If $M_A < M_B$ or $m_A > m_B$, then B is not a subset of A. Supposing that pruning fails, the intervals are subdivided and the subintervals tested in the same fashion, taking into account constraints of interval size and order. The process proceeds until a decision can be made.

Search is fast if pruning occurs soon, that is, at a high level in the hierarchy. Proving a match takes longer, but less than with straightforward shift-and-compare because misalignments are detected quickly. As an example of an application of such an algorithm, a desktop environment could visually highlight all files that contain a given piece of data, or that contain data duplicated elsewhere.

It is worth noting that search structures based on scalable statistics are not *optimal* for search. They are however scalable.

5. DISPLAY AND NAVIGATION

As content gets larger, more navigation must occur within metadata and less within data. Therefore, the problem is actually display and navigation of *metadata*. It is well known that hierarchical

strategies are the most effective for navigation over wide scales [17, 21, 23]. Scalable metadata support such strategies.

Content-based metadata can be used for display either directly (a picture of the waveform or spectrogram) or indirectly as ingredients for more sophisticated displays. A "waveform" descriptor scaled as a time-series of min/max pairs is sufficient to display that waveform. As long as there are as many pairs as pixels horizontally, the result is identical to plotting all samples of the original waveform, but much cheaper in terms of storage, transport and drawing. Similar comments may be made for the spectrum, for which a scaled series of spectrum slices is sufficient to obtain a high-quality display. The quality/cost ratio may be enhanced by associating several descriptors, or scalings of the same descriptors. For example an effect similar to an analog oscilloscope (greater brightness where the distribution of values is more dense) can be provided by associating histogram or mean/variance statistics to the basic min/max data used to plot the waveform. A spectrum descriptor with low frequency and/or temporal resolution can yield a "high-resolution" display by associating it with it a power descriptor (to enhance temporal resolution), an F_0 descriptor (to provide texture reflecting the harmonic structure), a harmonicity map (to restrict the harmonic texture to certain regions), etc..

Content-based metadata can also support displays based on search or classification algorithms. For example documents with certain content characteristics can be grouped or color-coded, etc.. Another simple example is content-based icons (waveform, spectrum) for documents, or the detection of duplicates.

6. SONIFICATION

Content-based metadata may also be used for browsing with auditory feedback, or as "earcons". A straightforward approach is to use a spectrum (or other spectral-shape descriptor) to design a time-varying filter that is then excited by a combination of noise and pulse trains (as defined by F_0 and harmonicity descriptors) [18]. Modulation descriptors can be used to synthesize beat, etc.. The result is an acoustic caricature of the content with two useful features. One that it reflects the *large scale* structure of a document. It is thus complementary to clip-type earcons based on selected portions of content. Another is that the mapping of content to sound is systematic and predictable, which is again not the case of human-edited clips.

7. SCALABLE METADATA IN MPEG-7

MPEG-7 is an initiative to develop and standardize tools for describing media [5, 6]. Standardization is crucial, as metadata are by nature shared between applications, and tend to have long lifetimes. Interoperability is paramount, and standardization is necessary for that. Unfortunately, scalability (as defined here) is implemented only partially in the MPEG-7 standard, and in different ways for video content and audio. For video the focus was on descriptors well optimized for space, but not necessarily scalable (scalability, where present, is usually limited in range and descriptor-dependent).

For audio, descriptors were layered over a scalable data structure named "Scalable Series" (originally ScaleTree) defined in XML Schema (Extensible Markup Language, http://www.w3.org/XML/). Audio descriptors are defined in an object-oriented fashion on the basis of this structure. One advantage of this approach is that development efforts to ensure scalability are concentrated in one place and their benefits shared across all descriptors. A second is that *rescaling* can be performed at the systems layer: the system needs to know only the scalable structure. A third is flexibility, as the "same" descriptor can be instantiated at different resolutions with different scaling operations, to cater for the widest range of needs. Applications can also use Scalable Series within their own specialized descriptors [19]

Scalability entails a cost in terms of interoperability, as it increases the range of resolutions in use for each type of metadata. This is unavoidable. Indeed, restricting flexibility would have the far worse effect of encouraging the proliferation of "different" descriptors with similar semantics. Scalability provides well-defined paths between resolutions, making them as interoperable as possible, even when different scaling operations are used. Scalability offers a good compromise between flexibility and interoperability.

Certain useful features are missing in MPEG-7 audio for the support of scalability. Contrary to MPEG-7 video, MPEG-7 audio initially placed little emphasis on optimization for storage. It was reasoned that efficient coding would be ensured either in a generic fashion at the systems level, or by future improvements of the design of the Scalable Series structure. Currently the underlying type (provided by XML Schema) is "float", and there is no obvious way to tell the generic binarization scheme of MPEG-7 what binary representation should be used to code each value. Lacking that, descriptions cannot be made as concise as they should.

A solution (within the constraints of XML Schema and the MPEG-7 Data Description Language) would be to define a quantization descriptor (QuantizationD), and use it to define a type for *series of quantized numbers*, which would then be used as the underlying data structure of Scalable Series. The quantization would be attached to each series, and used for interpretation by the application, and by the system to know how many bits to assign to each sample. This would enhance the compactness of all descriptors using Scalable Series.

Another missing feature is the *histogram*, currently not included among scaling methods (it too could be defined on the basis of a QuantizationD). Lacking also are mechanisms bridge the gap between with-document and accross-document descriptions in a scalable fashion, as well as scalable mechanisms to handle non-numerical metadata. It is hoped that development of MPEG-7 audio will fill these gaps, and that the idea of scalability will eventually spread to other areas of MPEG-7.

8. CONCLUSIONS

Scalability is a necessary property for metadata. Metadata can be made scalable in a variety of ways that cater to many applications. Scalable metadata are useful for tasks such as search, sonification and display, because of their inherently hierarchical nature, and because operations that ensure scalability also produce statistics that are valuable for these tasks. The notion of scalability has been used in the development of MPEG-7 audio. A basic data structure was introduced that confers scalability properties to content-based descriptors.

9. ACKNOWLEDGMENTS

This work was carried out with the support of the Cognitique programme of the French Ministry of Research and Technology.

10. REFERENCES

[1] A. M. Odlyzko, "The current state and likely evolution of the internet," in *Proc. Globecom '99, IEEE*, 1999, pp. 1869–1875.

[2] B. Schaller, "The origin, nature and implications of "moore's law" (http://mason.gmu.edu/∼rschalle /moorelaw.html)," 1996.

[3] M. Halem, F. Shaffer, N Palm, E. Salmon, S. Raghavan, and L. Kempster, "Technology assessment of high capacity data storage systems: can we avoid a data survivability crisis? (http://esdcd.gsfc.nasa.gov/ esdcd/whitepaper.data_survive.html)," 1999.

[4] K.G. Coffmann and A.M. Odlyzko, "Internet growth: Is there a "moore's law" for data traffic? (http://www. research.att.com/ amo/doc/networks.html)," 2000.

[5] Martinez, J. (2001), "Overview of the MPEG-7 standard," ISO/IEC JTC1/SC29/WG11 N4509.

[6] Pereira, F. (2001), "MPEG-7 Requirements Document V.16," ISO/IEC JTC1/SC29/WG11/N4510.

[7] "Mpeg-7 interoperability, conformance testing and profiling, version2," *ISO/IEC JTC 1/SC 29/WG 11, N4039*, 2000.

[8] A.K.Jain, M.N.Murty, and P.J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 30, pp. 265–321, 1999.

[9] V. Gaede and O. Günther, "Multidimensional access methods," *ACM Computing Surveys*, vol. 30, pp. 170–231, 1998.

[10] F. Crestani, M. Lalmas, C.J. Rijksbergen, and I. Campbell, ""is this document relevant?... probably": a survey of probabilistic models in information retrieval," *ACM Computing Surveys*, vol. 30, pp. 528–552, 1998.

[11] T. Palpanas, "Knowledge discovery in data warehouses," in *SIGMOD Record*, 2000, vol. 29, pp. 88–100.

[12] A. Berson and S.J. Smith, *Data warehousing, data mining and OLAP*, McGraw-Hill, New York, 1997.

[13] Sivakumaran, P., Fortuna, J., and Ariyaeeinia, A. M. (2001). "On the use of the Bayesian Information Criterion in Multiple Speaker Detection.", Proc. Eurospeech, 795-798.

[14] K. Kashino, G. Smith and H. Murase, "Time-series Active Search for Quick Retrieval of Audio and Video", IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP-99), Vol. VI, pp.2993-2996 (Mar. 1999)

[15] K. Melhorn, *Data structures and algorithms 3: Multidimensional searching and computational geometry*, Springer Verlag, Berlin, 1984.

[16] D. Gusfield, *Algorithms on strings, trees and sequences. Computer science and computational biology.*, Cambridge University Press, Cambridge, 1997.

[17] Y. Guiard, M. Beaudoin-Lafon, and D. Mottet, "Navigation as multiscale pointing: extending Fitt's model to very high-precision tasks," in *Proc. Computer-Human Interface (CHI)*, 1999, pp. 450–456.

[18] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequency cepstal coefficients and pitch frequency," in *Proc. IEEE ICASSP*, 2000, pp. 1299,1302.

[19] "Information technology multimedia content description interface part 4: Audio," *ISO/IEC CD 15938-4*, 2000.

[20] Chen, C., Gagaudis, G., and Rosin, P. (2000). "Similarity-based image browsing.", Proc. IFIP, 206-213.

[21] Cribbin, T., and Chen, C. (2001). "Visual-Spatial Exploration of Thematic Spaces: A Comparative Study of Three Visualisation Models.", Proc. Electronic Imaging 2001: Visual Data Exploration and Analysis VIII.

[22] Ghias, A., Logan, J., Chamberlin, D., and Smith, B. C. (1995). "Query by humming.", Proc. ACM Multimedia, 213-236.

[23] Herman, I., Melanon, G., and Marshall, M. S. (2000). "Graph visualization and navigation in information visualization: a survey," IEEE Trans. on visualization and computer graphics 6, 24-44.

[24] Khan, L. R., Shahbi, C., Alshayje, A., and Jiang, N. (1999). "Improving the performance of audio-based similarity queries with clustering.", Proc. ACM Workshop on Multimedia Intelligent Storage and Retrieval Management, Orlando.

[25] Luettgen, M. R., and Willsky, A. S. (1995). "Likelihood calculation for a class of multiscale stochastic models, with application to texture discrimination," IEEE Trans. Image Proc. 4, 194-207.

[26] Spence, C., and Parra, L. (2000). "Hierarchical Image Probability (HIP) Models.", Proc. Advances in Neural Information Processing Systems, 848-854.

[27] Subramanya, S. R., Youssef, A., Narahara, B., and Simha, R. (1997). "Transform-based indexing for multimedia systems.", Proc. IEEE Int'l Conference on Multimedia Systems, Ottawa.

[28] de Cheveigné, A., and Peeters, G. (2000), "Core set of audio signal descriptors," ISO/IEC JTC1/SC29/WG11, MPEG00/m5885 technical report.

[29] de Cheveigné, A., and Peeters, G. (1999), "Scale tree," ISO/IEC JTC1/SC29/WG11, MPEG99/m5076 technical report.

[30] de Cheveigné, A. (1999), "Scale tree update," ISO/IEC JTC1/SC29/WG11, MPEG99/m5443 technical report.

[31] de Cheveigné, A., and Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. 111, 1917-1930.