The Hifi of the Future : Toward new Modes of Music-Ing

François Pachet

Sony Computer Science Laboratories - Paris

Abstract

His paper argues that the evolution of Hifi systems is undertaking major changes, due to the possibility of creating and manipulating metadata for large music collections. We describe here an ongoing project about the Hifi of the future, and present several innovative applications dealing with musical metadata in an integrated way. Three hitherto separated musical activities: accessing, performing and learning can be united in a single environment, allowing many new modes of "making" music.

Keywords: Hifi, Music, metadata.

1 Introduction : The Diversity of Musical Activities

The recent collapse of the traditional music industry (strong decrease in CD sales) is a symptom of a crisis which is not simply a commercial one. Indeed, this paper argues that it is not only the CD sales which have come to an end, but also the very traditional *usages* of music, in particular music listening which have now to be reconsidered. We stress on the fact here that the technologies referred to as "content technologies", in particular metadata, offer radically new ways of envisioning music enjoyment in the large. By metadata we mean information associated to music, either manually or automatically, that enhance the actual music information. If metadata has been a buzz word these last years, it is only recently that powerful metadata extraction mechanisms have been designed, allowing designers to think about new, operational, music applications.

More precisely, this papers argues that three types of music relationships can, and must, be brought together, to provide new music environments for the future:

- *Music access*. By access we mean both the traditional music lover behaviour of listening to a piece of music and also the new access modes permitted by the digitisation of large music collections, through music browsers of various kinds. More generally, the shift here is to go from the access to a music piece (or work), to access to a *collection*. This collection can be personal or can be editorial (e.g. the Vivendi Universal music catalogue).
- *Music Interactions.* Interacting with music is not a new concept: the volume button on amplifiers offers listeners a simple but efficient way to interact with music – by making it louder. However, today, thanks to the possibility to extract meaningful information, but also to synthesize in real time, more "interesting" modes of interaction with music can be invented.
- *Learning and Education.* Learning music is traditionally performed in specific locations , such as conservatories. Today, many software can help drastically the learner, either by assisting existing, traditional pedagogical curriculum, or by providing new forms of learning schemes, e.g. through edutainment.

The goal of this paper is to briefly sketch a vision where these three different modes of music-ing (accessing, interaction, learning) can be put together to create something which is more than the sum of its parts: the Hifi of the future. The vision is embodied in the form of various interrelated applications, to be seen as a prototype platform for the Hifi of the Future.

2 The Evolution of Hifi Systems: From Buttons to Exploration

We propose the idea of exploratory listening environments, as a natural evolution in Hifi systems, and more precisely in a history of "musical controls". We first sketch a brief history of musical controls, and then introduce the notion of semantic-preserving musical exploratory environment.

2.1 History of Musical Controls

Each technological advance has brought with it new forms of controls. The origins of listening machines with mass-produced musical materials may be traced back to the Phonograph, invented by Thomas Edison in 1878, which used tin foil cylinders, and shortly after the Gramophone, invented by Berliner in 1888, which used flat disks. In these devices, there was no control intentionally given to the user (se, e.g. Read & Welch, 1976). There was, however, an unintentional control in the Gramophone in that the horn could be *turned* around, thereby influencing the directivity of the sound source. Electricity soon began to be used for listening devices, both with radio and with new electrically recorded disk players in the 20s. The use of electricity also introduced new controls: the *volume* button and the *treble/bass* button. Juke-boxes were introduced in 1927, allowing listeners to *select* explicitly music titles from a given catalogue of disks, using various sorts of push buttons. The next big technological advance was the invention of binaural (stereo) recording method in 1931. The corresponding control was the *panoramic* button allowing to control the amount of signal in one loudspeaker or the other. Finally, digital format for audio introduced more controls, e.g. on the equalization of sound. In all these cases,

technological advances were followed by the introduction of “technical” controls, i.e. controls operating directly on the technology (see Figure 1).

Figure 1. A Phonograph (Edison, 1978, left); a Gramophone (Berliner, 1988, middle), a Rock-Ola 120-selection Juke-Box, and a Mini disk player (Sony, 1997, right). Advances in technology do not necessary imply more intelligent user control.

2.2 A Matter of Semantics

The very notion of musical control raises the issue of *semantics*. The issue of musical semantics - does music have meaning ? - has been long debated by musicologists, leading to different theories, which usually paralleled the theories of semantics for languages. One of the main distinction made by theorists is the opposition between so-called “referentialists” and “absolutists”. Referentialists claim that musical meaning comes from actual references of musical forms to outside objects, i.e. music means something which is external to music itself. For instance, a particular scale in Indian music may have a reference to a particular human mood. Absolutists, e.g. Strawinsky, claim on the contrary that the meaning of music, if any, lies in music itself, i.e. in the relations entertained by musical forms together. Although these two viewpoints are not necessarily exclusive, as noted by Meyer (Meyer, 1956), they leave open much of the question of meaning. Eugene Narmour elaborated a much more precise theory of musical meaning based on the psychological notion of expectation (Narmour, 1992). In this theory, meaning occurs only when musical expectation are deceived. On the other hand, Rosen argues (Rosen, 1994) that the

responsibility of preserving the meaning of a musical piece lies only in the performer itself, who has to choose carefully among a infinite set of possible interpretations which one is closest to the one “intended” by the composer.

Without committing to one particular theory of musical meaning, we note that meaning - whatever it means - has to do with *choosing* among a set of interpretations the “right one” or the “right ones”, e.g. those intended by the composer. A second remark is that the controls given by the history of sound recording technology have never had any concern about musical semantics: what does it mean to raise the sound level of a record ? to shift the signal to the left loudspeaker ? to increase the bass frequency ? Are the intentions of the composers, or even of the sound engineers, preserved in any way ?

From this remark, we suggest that the new “interesting” musical controls of the Hifi of the Future should preserve some sort of semantics of the musical material, i.e. preserve intentions, whenever possible. We argue that more meaningful controls, in the context of modern digital multimedia technology, amount to shifting from traditional button-based technology to musical exploration spaces.

2.3 Music Interactivity

As we have seen, technological buttons bear no semantics, because they are directly grounded on the technology, without any model of the music being played. But what could be such a model ?

Interesting approaches in musical interactivity are the music notation systems, in the context of annotation of music documents, as in the works of Lepain (1998), or in the Acousmograph system (INA-GRM). In these systems, the primary issue addressed is not music listening per se, but rather music *notation*, i.e. how to represent graphically

a musical document (the document itself or the perception of the document), or how to infer a model of the music which can be noted or represented graphically.

Another answer may be found in the notion of *open form*, initially developed in literature (Eco, 1962), which has had much impact on music theory and composition (Stockhausen, Boulez). The idea of musical open form is that the composer does not create a ready-to-use score, but rather a set of potential performances, which can be seen as a *model* of scores, as explained by (Eckel, 1997): "Music is not any longer conceived in form of finite units but in terms of models capable of producing a potentially infinite number of variants of a particular family of musical ideas". The selection or instantiation of the actual score to be played is delegated to the performer. In recent incarnations of open form, it is the listener himself who instantiates the model, as for instance in the *Cave* (Cruz-Neira et al., 1993) or *CyberStage* (Eckel, 1997). In these cases, the user is immersed in a realistic virtual environment, and has the control on his position and movement in a virtual world. His movements are translated into variations in the musical material being heard. These approaches may be considered as radical, in the sense that the user has a great deal of responsibility in making the music. However, the issue of semantics is not directly addressed, since the model in principle is under-designed, i.e. all possible explorations are always "licit", whatever they may be. In this respect, there is a strong relation between open form virtual environments and programming languages for music composition, such as *OpenMusic* (Assayag et al., 1997). In these approaches indeed, the goal is to propose the user to explore spaces with as much freedom as possible, and not constrain the user in specific areas.

2.4 Active Listening

Active Listening refers to the idea that listeners can be given some degree of control on the music they listen to, that gives the possibility of proposing different musical perceptions on a piece of music, by opposition to traditional listening, in which the musical media is played passively by some neutral device. The objective is both to increase the musical comfort of listeners, and, when possible, to provide listeners with smoother paths to new music (music they do not know, or do not like). Active listening is thus related to the notion of open form outlined above but differs by two important aspects: 1) we seek to create listening environments for existing music repertoires, rather than creating environments for composition or free musical exploration and 2) we aim at creating environments in which the variations always preserve the original semantics of the music, at least when this semantics can be defined precisely. For us, the issue if therefore not to introduce yet another technological button in the interface of the listening device, but rather to design buttons that “make sense”, thereby breaking the long tradition of technological buttons initiated by Edison.

What “sense”, what “meaning” are we talking about ? How can music controls be designed to trigger semantic preserving actions ? The answer stems from the new landscape of music recording created by digital multimedia, sketched in the next section. The vision of the Hifi of the Future sketeched in section 4 stems directly from this landscape.

3 The New Facts of Multimedia

Digitalization of multimedia data has a number of technical advantages which are well known today: better sound quality, better compression, lossless copy, etc. The

aim of this chapter is to show that digitalization of multimedia data also induce - even in a still potential form - a number of revolutions in the way music may be accessed and listened to by end users. We will outline three of these revolutions, which form the basis of our argumentation, focusing on the paradigmatic shifts they convey, rather than on technical aspects.

3.1 Structured Audio: Home as a Reconstruction Machine

The idea of structured audio has initially been devised to allow better compression of high quality audio. Standardization efforts like the Mpeg-4 project embody this idea, and try to make it practical on a large scale (see, e.g. the Machine listening Group of the Media lab, Sheirer et al., 1998).

The idea is simple: instead of transmitting a ready-to-listen sound, only a description of how to make the sound is transmitted. The actual sound is reconstructed at home, or at the listener's location, provided of course he/she has the right software to process this reconstruction properly. Structured audio actually extends this basic idea to include fully-fledged *scene descriptions*, that is, not only descriptions of individual sounds, but description of groups of sounds playing together to make up a piece of music. The actual technical details of scene description also include all what is needed to reconstruct a sound or piece of music rightfully, e.g. effects, adaptation to the local sound reproduction system, and so forth.

Such a notion of scene description opens up new doors for meaningful controls. Indeed, since the music is delivered as a "kit", lots of possibilities can be imagined to influence the way the kit is actually built, according to user preferences. Of course, these variations around how the kit should be assembled have to be "coherent", which are precisely the matter of our work.

3.2 Meta-data and All That Jazz

The fact that musical data is now produced, coded and transmitted in a digital form has numerous and well-known advantages: better sound quality, possibility of lossless transmission and copying (thereby raising new copyright problems). An important non technical consequence is the possibility to encode not only the music itself - the digitalized sound - but also any sort of symbolic information. Such symbolic information may be used to code and transmit data on the music itself, so-called information on content, meta-data or also “bits about bits”.

Why would one want to transmit such meta-data ? The interests are obvious in the context of document indexing. If musical data is accompanied with corresponding adequate descriptions, digital catalogues can then be accessed using sophisticated query systems. Current standardization efforts like Mpeg-7 embody this idea (MPEG7, 1998), and try to define standards for describing meta-data for all sorts of multimedia documents. MPEG-7 aims for instance at making the web more searchable for multimedia content than it is today, make large content archives accessible to the public.

Here again, we would like to emphasize the conceptual rather than the technical aspects of this paradigm shift: meta-data opens also doors for imagining new listening systems in which the user may access data in a drastically different way. Instead of being a passive, neutral support, music becomes an active, self-documented knowledge base. Again, what kind of listening devices can be imagined that exploit this information ?

3.3 Size of Digital Catalogues

Digitalization of multimedia data has yet another consequence: the availability of huge catalogues of multimedia data to users. In the case of music, there is, here also, a conceptual shift which has nothing to do with the technology of large databases. The main issue raised by this technological advance is how to access huge catalogues of music, not from a technical viewpoint, but from a user's viewpoint. Recall the juke box, invented in the late 20s: a typical juke box would contain about 120 titles, which is the size of an average user's discotheque. Browsing through all the titles was probably part of the pleasure, and selection could be made just like at home: by choosing one item out of a collection of items, which at least the user has seen once. Now a typical catalogue of a major company is about 500.000 items. What happens when the collection to select from is such a catalogue ? Even more terrifying, what happens if all the recorded titles become available through networks to users at home ? Estimating the total number of all recorded music is difficult, but it can be approximated to about 10 million titles. The figure can be probably doubled to include non Western music. Every month, about 4000 new CDs are issued on the market. It is clearly impossible to apply usual techniques of music selection in this new context. What does it mean to "look for" a title when the mass of titles is so huge ?

4 The Hifi of the Future

The Hifi of the Future is the name of an ongoing project trying to answer to questions raised above. This project is conducted at Sony Computer Science Laboratories in Paris and has been running since 1997, with the development of several prototypes dealing with musical exploration: *ProgramBuilder* (Pachet et Roy, 1999), *PathFinder* (Pachet et al., 2000), *PersonalRadio* (Pachet, 2003), *MusicBrowser* (Pachet et al,

2004). These tools have been developed and used in various prototyping environments, such as the European funded Cuidado project (Vinet et al, 2002) and the ongoing Semantic-Hifi project (started in December 2003). Experiments with users have been conducted at University of Bologna, as well as Cité des Sciences, notably during the Villette numérique festival(www.villette-numerique.com/2004/pages/splash.php). Based on this accumulated experience we have come up with a vision that only metadata-enhanced musical applications can provide novel music usage modes. More precisely we propose three working hypothesis for building an environment that integrates smoothly all the techniques sketched above in an integrated manner, so as to propose novel applications that can expand the possibilities of music access.

4.1 Three Hypothesis

- Integration of activities in a *single environment*. The different applications envisaged will necessarily share many information, data, metadata and also software components; It is therefore crucial that they can communicate with each other smoothly.
- Need for managing efficiently *large databases*. Metadata is interesting, by definition, only for managing large databases, which in turn creates issues of efficiency. Compilers which create efficient Sql queries are mandatory to create systems useable by non professionals.
- Need for *vertical languages* to develop these new systems. The development of a content-based music application requires the handling of many different layers of software development, from the design of audio acoustic descriptors to the development of graphical interfaces (Matlab, C++, Sql, Php, Java, etc.). Managing these different levels and their interconnections is time consuming

and acrobatic. Vertical languages reduce the difficulty by packaging vertically services, thereby freeing the developer to handle manually all these levels.

4.2 The MCM Framework

To implement the hypothesis proposed above, we have developed an object-oriented framework (in the sense of (Fayad et al. 1999)) called *MCM* (standing for Multimedia Content Management). This framework contains all the important services needed to build content-based music applications, from the design of perceptive descriptors (using the EDS system (Zils & Pachet, 2004)) to the creation of specific ontologies such as genre (Laburthe et al. 2003) and the creation of user interfaces. This paper does not give details on MCM, which can be found on other publications of the team. We will here only provide a small exemplar list of applications built with MCM making use of these services.

4.3 Examples of new applications

In this section, we describe some typical applications built with MCM. This list is by no means intended to be complete, and is given as an indication of the types of applications that can populate the Hifi of the Future.

4.3.1 EyeTune

EyeTune is an application of gesture recognition techniques for the management of large music collections (borrowing its name from the popular EyeToy game (Kushner, 2004). EyeTune allows users to perform basic and simple controls such as volume control, but also to select music using high-level metadata, such as genre, country, artist, etc. (see Figure 2). The simplicity of the application (no need for any mouse or keyboard) makes it particularly adapted to the family context. Also, control can be

effected by more than one person at the same time. Specific gestures are being developed to allow controls adapted to music access, such as tempo tapping for query by tempo, or line drawing for creating personalized playlists with given properties (e.g. ascending tempo, etc.).

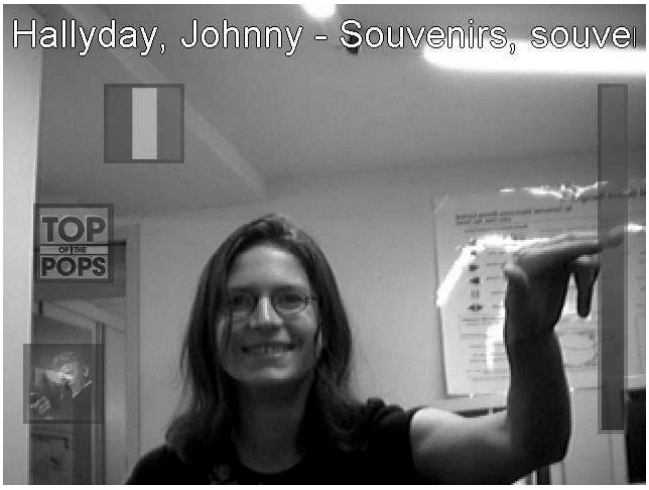

Figure 2. A user of the EyeTune application, selecting music by simple metadata through gesture recognition techniques. Here, controlling the volume level.

4.3.2 Advanced MusicBrowser

The Advanced MusicBrowser (see Figure 3) allows users to browse within a collection of music titles. The main difference with the standard music browser described e.g. in (Pachet et al., 2004) is the use of the reflective capacities offered by MCM. Indeed, not only simple queries can be issued, e.g. about songs (such as find the songs whose country is "UK"), but arbitrarily complex queries can be built to find more fine-grained information such as "find the country with artists of the greatest number of different styles", or "Find the countries closest to France in terms of having songs with the same voice and tempo characteristics". Additionally, Artificial intelligence techniques allows to *infer* properties from the analysis of a given catalogue. For instance, the user can select an item (song, artist, country, etc.) and ask the system to find "interesting" properties characterizing the item. For instance, a

song can be interesting because it is the only representative of a given category (e.g. genre, voice type), or a given combination of categories (i.e. the only French song with electric guitar), etc.

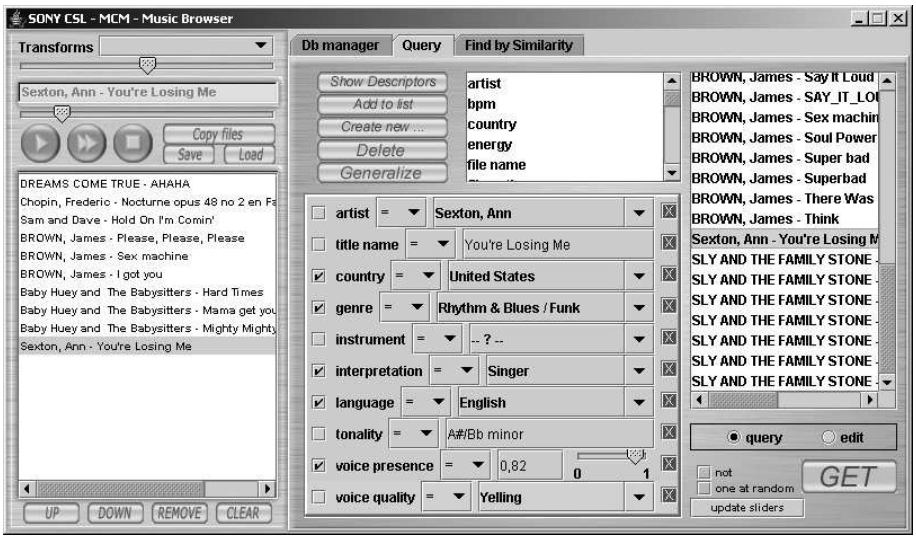

Figure 3. The Advanced Music Browser provides sophisticated ways of browsing through a large music collection using arbitrary combinations of descriptors and metadata.

4.3.3 Baby Browser

The BabyBrowser is designed for 4 to 5 years old children which are in the reading learning stage. The system allows to browse video clips (e.g. musical) by typing in artist and song information. The tool proposes a simple field text with an associated word completion mechanism and a speech synthesizer. The child can see the list of available artists on ta list, but cannot click on them, and has to type characteristic letters to select an artist. Each letter typed triggers the speech synthesizer, so the child can readily understand the effect of letter associations. When an artist is selected, the same process occurs for selecting the title. Eventually, the corresponding video clip is played (see Figure 4). Current studies attempt to show how reading skills can improve by using the tool (in conjunction with traditional learning methods). We strongly

believe that this type of edutainment software can be very efficient thanks to the popularity of music and videos, which create strong and natural motivations to learn.

Figure 4. A child using the BabyBrowser. Selection of video clips is performed using a mix of textual input and speech synthesizer. Although the goal is for the child to watch video clips, the hidden goals, for the system, is to teach how to read (and write).

4.3.4 SongSampler

The SongSampler is a typical application mixing together the world of interaction and the world of access in an original way. Basically the idea is to create on-the-fly music instruments (more exactly samplers) from the analysis of arbitrary songs. The user listens to a song and can at any moment decide to create a corresponding "instrument". This instrument will automatically segment the song and identify "interesting" segments, and create a sampler from them. The user can then play with the sampler his own music, but with the sounds the original song (Aucouturier & Pachet, 2004).

4.3.5 Musaicing

The idea of musaicing is a transposition of the notion of image mosaicing to the world of audio (Zils & Pachet, 2001). Musaicing makes intensive use of large databases of audio samples, and allows user to create music without having to perform the tedious

and difficult task of listening and selecting individual samples. Musaicing consists in creating automatically large databases of samples by segmenting existing songs (see Figure 5). Then metadata is computed for each sample to describe it in terms of perceptive parameters (such as timbre, percussivity, energy, pitchness, etc.). Finally the user can express high-level *constraints* to specify the structure and nature of a target sequence of samples. Constraints can be of various types, such as *continuity* (produce a sequence of samples which are continuous, timbre-wise), *distribution* (select a percussive sample every beat with tempo = 120) or *cardinality* (include at least 40% of samples which come from a Beatles song), or any combinations of these.

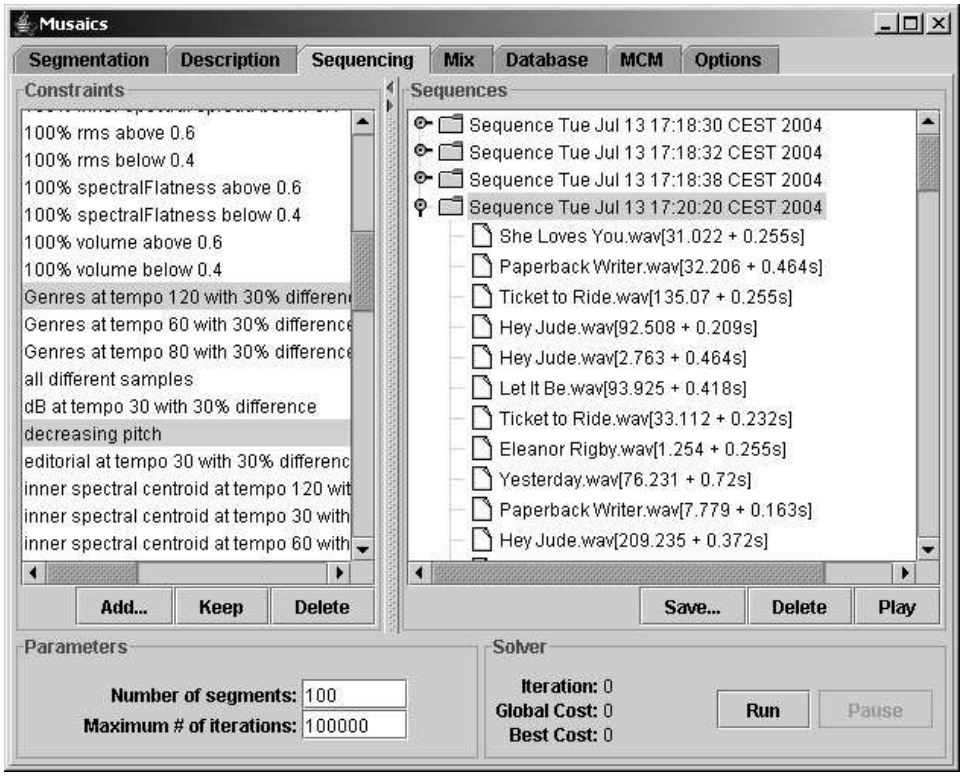

Figure 5. The musaicing system allows user to create sequences of sounds (i.e. music) from large collections of samples obtained automatically from the segmentation of existing songs.

5 Conclusion

The new possibilities offered by the systematic digitisation of popular culture, together with the power of metadata techniques makes it possible to envisage today new modes of music enjoyment. A complete list of such modes is yet to be determined, but we propose here several cases of music applications based on putting together listening, access performing and learning. The development of these modes require non trivial manipulations of a large set of computer technologies, ranging from descriptor extraction to gesture recognition. These manipulation are greatly simplified by the use of dedicated framework, such as MCM. But these applications are only a starting point and it is hoped that many more ideas can be found to create novel modes of music access and enjoyment that make use meaningfully of these huge quantities of music now being offered to us.

6 References

Assayag G., Agon C., Fineberg, J., Hanappe P., "An Object Oriented Visual Environment For Musical Composition", *Proceedings of the International Computer Music Conference*, pp. 364-367, Thessaloniki, 1997.

Aucouturier, J.-J., Pachet, F. (2004) From Sound Sampling to Song Samplers, Proc. of ISMIR 2004, Barcelona, November.

Cruz-Neira, C., Leight, J., Papka, M., Barnes, C., Cohen, S.M., Das, S., Engelmann, R., Hudson, R., Roy, T., Siegel, L., Vasilakis, C., DeFanti, T.A., Sandin, D.J., "Scientists in Wonderland: a Report on Visualization Applications in the CAVE Virtual Reality Environment", Proc. IEEE Symp. on Research Frontiers in V.R., pp. 59-66, 1993.

Eco, U. Opera Aperta. Bompiani (Milan), 1962.

Eckel G., "Exploring Musical Space by Means of Virtual Architecture", *Proceedings of the 8th International Symposium on Electronic Art*, School of the Art Institute of Chicago, 1997.

Kushner, D. (2004) Computing gets Physical. Technology Review, July/August 2004.

La Burthe, A., Pachet, F. and Aucouturier, JJ Editorial Metadata in the Cuidado Music Browser: Between Universalism and Autism. Proceedings of the WedelMusic Conference, 2003

Lepain, P. Ecoute interactive des documents musicaux numériques, in *Recherches et Applications en Informatique Musicale*, Chemillier & Pachet Eds, Hermes, Paris, 1998.

Meyer, L. *Emotions and meaning in Music*, University of Chicago Press, 1956.

MPEG7 Requirements Group, "MPEG-7 Requirements Document", Doc. ISO/MPEG N2461, MPEG Atlantic City Meeting, October 1998.

Narmour, E. *The analysis and cognition of melodic complexity*. University of Chicago Press, 1992.

Pachet, F. (1999) Constraints for Multimedia Applications. Proceedings of PACLP 1999, The Practical Application Company, London, March.

Pachet, F., Roy, P. and Cazaly, D. (2000) A Combinatorial Approach to Content-Based Music Selection. IEEE Multimedia, 7(1):44-51 March.

Pachet, F. Laburthe, A., Aucouturier, J.-J. (2004) Popular Music Access: The Sony Music Browser. Journal of American Society for Information Science, Special issue on music metadata, Downie, S. Editor.

Pachet, F. and Delerue, O. (2000) On-The-Fly Multi-Track Mixing. Proceedings of AES 109th Convention, Los Angeles. Audio Engineering Society.

Read, Oliver and Walter Welch. From Tin Foil to Stereo: Evolution of the Phonograph. Indianapolis, 1959, 2nd edition 1976.

Rosen, C. The Frontiers of Meaning: Three Informal Lectures on Music, Hill & Wang Pub, 1994

Sheirer, E. Väänänen, R. Huopaniemi, J. "AudioBiffs: The MPEG-4 Standard for Effects Processing", *Proceedings of the first Digital Audio Effects Workshop*, Barcelona, pp. 159-167, November 1998.

Vinet, H., Herrera, P., Pachet, F. (2002) The Cuidado Project: New Applications Based on Audio and Music Content Description. In ICMA, editor, Proceedings of ICMC, pages 450-454, September 2002. ICMA.

Zils, A. and Pachet, F (2001) Musical Mosaicing. Proceedings of Digital Audio Effects Conference, DAFX'01, December, University of Limerick.