The Representation Levels of Music Information

Hugues Vinet

IRCAM
1, place Igor Stravinsky
F-75004 PARIS - FRANCE
hugues.vinet@ircam.fr
http://www.ircam.fr

Abstract. The purpose of this article is to characterize the various kinds and specificities of music representations in technical systems. It shows that an appropriate division derived from existing applications relies in four main types, which are defined as the *physical*, *signal*, *symbolic* and *knowledge* levels. This fair simple and straightforward division provides a powerful grid for analyzing all kinds of musical applications, up to the ones resulting from the most recent research advances. Moreover, it is particularly adapted to exhibiting most current scientific issues in music technology as problems of *conversion between various representation levels*. The effectiveness of these concepts is then illustrated through an overview of existing applications functionalities, in particular from examples of recent research performed at IRCAM.

1 Introduction

The growing importance of the music industry as a key economic sector combined with the current convergence of computer, audiovisual and telecommunication technologies, yields rapid developments in music technologies. These technical evolutions have an impact at all levels of the production chain (production, publishing, dissemination and consumption), and bring new modalities of presentation, access to, and manipulation of the music material. As a result, there is an unprecedented variety of applications resulting from the music technology industry and research. Examples of such applications include score editors, MIDI and audio sequencers, real time DSP modules, virtual instruments based on physical modeling, computer-aided composition environments, 3D audio rendering systems, title databases with content-based browsing features, etc.

A question then arises□ given the various approaches to the music phenomenon developed in these applications, is it possible to derive a global view, which integrates all kinds of associated representations in a single, unified scheme? In the context of this first issue of CMMR and its focus on music modeling issues, the purpose of this article is to answer this question by characterizing the specificities of music representations in technical systems. Therefore, the proposed approach relies on the identification of a limited number of well-defined representation types, called *Representation Levels,* or *RLs*, for reasons to be further developed, and to analyze existing applications through this RL grid. The effectiveness of these concepts will be then illustrated through an overview of existing applications features, in particular from recent research performed at IRCAM.

2 Definitions and Properties of the Representation Levels

There are multiple ways of representing music information in technical systems, and in particular with computers. Such representations are chosen according to relevant viewpoints on the music content in order to match the target system functions. The term "Representation" refers here to the way information is represented internally in the system, i.e. essentially data structures. This article, in its aforementioned scope, does not handle another complementary aspect of representations in applications, related to man-machine interfaces, i.e. the way internal data are mediated to the user and, inversely, the way he can access them for manipulation. These issues are handled elsewhere in the context of man-machine interfaces for music production [23].

2.1 Music Representation Types

The conception of various kinds of music representation types is motivated by the recent history of music technology. There has been for several decades two main distinct and complementary ways of representing music content in technical systems□

- audio *signal representations*, resulting from the recording of sound sources or from direct electronic synthesis,
- *symbolic representations*, i.e. representations of discrete musical events such as notes, chords, rhythms, etc.

This distinction has been effective since the very beginning of computer music and is respectively exemplified by the pioneering works, almost simultaneous in the 1950s, of Max Matthews for the first digital music syntheses [10] and Lejaren Hiller for music compositional algorithms[7]. It is still true with current commercial music applications such as sequencers, in which digital audio and MIDI formats coexist.

Fundamental differences between both representations can be expressed as follows:

- the symbolic representation is *content-aware* and describes events in relation to formalized concepts of music (music theory), whereas the signal representation is a blind[1], content-unaware representation, thus adapted to transmit any, non-musical kind of sound, and even non-audible signals[2].
- even digitized through sampling, the signal representation appears as a continuous flow of information, both in time and amplitude, whereas the sym-

[1] One should rather say *deaf* in this context, but languages provide few auditory metaphors…

[2] For instance, a sinusoidal function at a 1Hz frequency.

bolic representation accounts for discrete events, both in time and in possible event states (e.g. pitch scales). Low bandwidth control parameters, such as MIDI continuous controllers, are also part of this category.

It should also be noted that despite various existing methods for coding audio signals, be them analog or digital, even in compressed form, they all refer to and enable the reconstruction of the same representation of audio signals as amplitude functions of time. On the contrary, symbolic representations gather a variety of descriptive approaches, including control-based information such as in the MIDI and General MIDI standards, score descriptions used in score editors, or more sophisticated, object-oriented musical structures found in computer-aided composition environments. For instance, the OpenMusic environment, developed at IRCAM, is a visual programming environment which enables, as illustrated in Figure□1, to design processing functions of symbolic information [2].

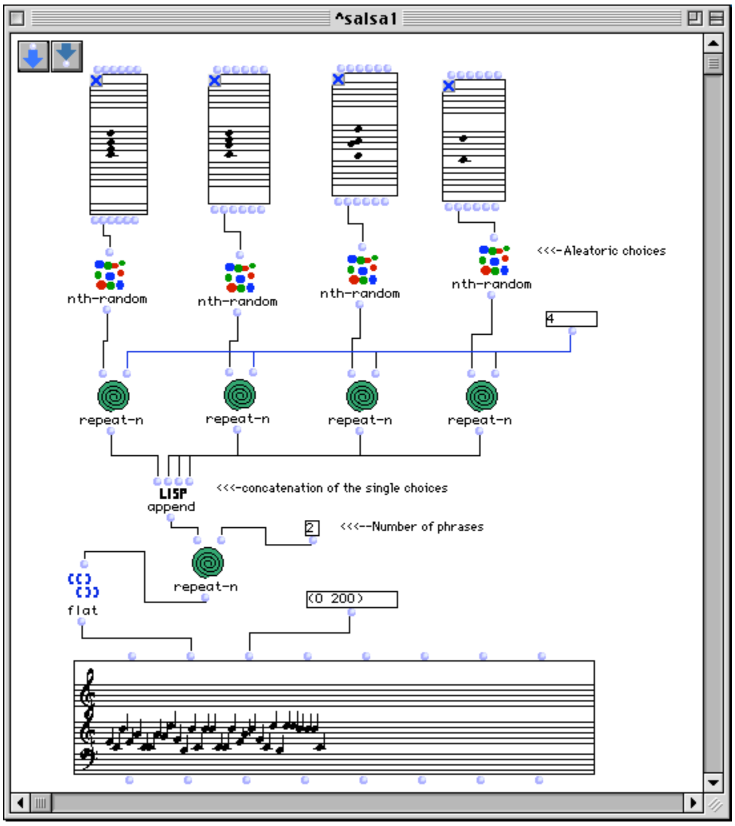

Fig. 1. Patch example in OpenMusic. The Input materials (chords) are positioned on top of the window, and the produced result is displayed in a notation editor at the bottom.

However, these two kinds of representations are not sufficient for characterizing all aspects of music contents found in existing technology. In order to take into account recent advances in musical applications, it is necessary to integrate two other kinds of representations, hereinafter defined as *physical* and *knowledge* representations.

Physical representations result from physical descriptions of musical phenomena, in particular through acoustic models. As one of their specificities, these representations account for *spatial characteristics* of sound objects and scenes, in terms of geometrical descriptions, but also include other physical properties, e.g. mass, elasticity and viscosity. The introduction of physical representations is first motivated by the growing importance of physical models of sound sources for audio synthesis, in terms of excitation and oscillation, but also through new concerns related to radiation synthesis[13], i.e. the reproduction of directivity patterns through multi-excitator systems. Second, these representations are also necessary for accounting for new spatialization applications, including. 3D audio simulations, which do not rely any more only on fixed multichannel reproduction setups (stereo, 5.1, etc.), but also include geometrical descriptions of sound scenes. Finally, advanced applications in the context of virtual and augmented reality, or new instruments, require the explicit modeling of gestural control information resulting from motion or gestural capture systems.

Knowledge representations provide structured formalizations of useful knowledge on musical objects for specific applications, such as music multimedia libraries. These representations rely on structures extracted from language as a conceptual basis for describing musical phenomena, whereas physical and signal representations rely on mathematical formalisms. Knowledge representations are thus essentially made of textual descriptions, and are adapted in particular to providing qualitative descriptions, which the other representations kinds do not enable. They have been developed in particular in the context of digital libraries and are currently the basis of numerous developments in the field of music information retrieval[24]. Unlike other music representation categories, there is no musical specificity of such representations, which can be considered for any knowledge area. Only their content, in the form of various knowledge representation structures, is to be designed specifically for particular music applications.

2.2 From Representation Types to Levels

In a scale which goes, from bottom up, from concrete to abstract descriptions, the four representation types which have been introduced can be ordered as follows:

Table 1. Ordering of Music Representation Types

Knowledge Representations
Symbolic Representations
Signal Representations
Physical Representations

An analogy could be found between the resulting levels and the various stages of musical information processing by the brains, from the auditory system physiology up to the highest cognitive levels. In this analogy, which is valid up to a certain extent, the signal level would correspond to the binaural signals, i.e. the acoustic pressure signals at the level of both eardrums, which characterize information inputted into the auditory system. Subsequently, the physical layer would correspond to the spatial body configuration in terms of position, orientation, and morphology, which intervene as the main factors of transfer functions between the 3D acoustic space and the binaural signals. The Symbolic level would be associated to the listener's knowledge of music material in reference to music theories or cultures, including the way he associates auditory percepts to categories issued from these theories (e.g. pitch quantization into scales). At the highest level, the knowledge representation is by definition associated to appropriate language structures for describing musical phenomena. This mapping is illustrated in figure□.

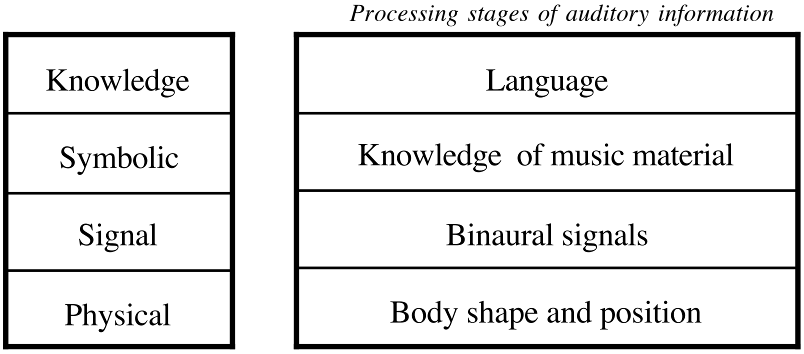

Fig. 2. Hypothetic mapping between representation levels and processing stages of auditory information.

However, this analogy presents some limitations, due to its inadequacy to reflect the structural complexity of the auditory system, in particular at the intermediary levels associated to the spectro-temporal coding of auditory information, and to model the way higher-level information is structured, which remains fairly unknown.

Since these concepts are intended to modeling music representations in technical systems, and do not pretend to model music cognition, a more objective viewpoint for justifying this ordered organization can actually be found in information theory. Increasing levels are obviously associated to a decreasing information quantity and the information reduction operated between successive layers takes the following specific forms□

- *from physical to signal levels*, the information reduction is essentially of *spatial nature* : the physical layer can be characterized by the acoustic pressure as a function of space and time, whereas the signal level is a function of time, which corresponds to the acoustic pressure signal captured by a microphone at a given space position.
- *from signal to symbolic levels*, the information reduction is associated to a *double digitization*, which concerns both the *time axis* and the *value ranges*

taken by analyzed variables. For instance, in typical signal to MIDI applications, the fundamental frequency is extracted from audio signals as a low bandwidth, slow variation signal, then it is again quantized over time into note events, and the frequency values are mapped into a discrete semitone pitch scale.

- *from signal and symbolic to knowledge representations*, the information reduction is generally the result on one side of a *global temporal integration* at a more or less global time scale, up to the whole piece duration, and on the other side the *projection of appropriate data value combinations to discrete categories*. Knowledge representations actually describe global characteristics of the music material, including objective descriptions such as the piece name, the music genre, the performing artists, the instruments played, as well as qualitative statements related to performance, sound quality, etc. In specific cases, these characteristics can be inferred from combinations of signal and symbolic information. For instance, if “rock” and “baroque” are relevant genre categories, one can consider inferring them from a combination of characteristic rhythmic and harmonic patterns extracted from symbolic information, and of spectral distributions associated to the characteristic timbres of associated instrument groups. It is also worth mentioning that characterizing qualitative aspects of performance will require information present in audio recordings, even if experiments have shown that interesting features can already be extracted from MIDI performance recordings , only through onset note positions and velocities[4].

2.3 Mapping of Existing Music Standards into the RL scale

In order to further specify the RL definitions, let us examine how various music data standards map to the RL scale.

Audio signal formats. Among mono- or stereophonic signals, the most complete representation, i.e. the one with the biggest information quantity, corresponds to analog signals. Digital audio signal are positioned higher in the RL scale, since they result from a double digitization, in time and amplitude, whose translation in terms of information reduction corresponds to a limitation of bandwidth and dynamics (or signal on noise ratio). The various digital audio formats (e.g. AES-EBU) are ranked in the RL scale according to their information quantity per time unit, i.e. as a function of sampling frequency and entropy of single word coding. Multichannel audio coding formats, such as MADI or ADAT, enable the coding of N independent signals, and it is easy to show that their entropy adds a constant $\log_2 N$ offset to the one of single coded signals. Audio compression formats, such as the ones found in MPEG1 (including the popular mp3 format), MPEG2 and MPEG-AAC, operate a significant information reduction (typically 10 times for mp3) without audible effects in broad sound classes, through the integration of psychoacoustic masking effects. This case illustrates once more the difficulty of setting up a straightforward mapping between the RLs and processing stages in the auditory system.

Symbolic representation formats. MIDI, the most widely spread standard for symbolic representations, combines the coding of various information types, including note events and continuous controllers, through various channels. Its low bit rate and value representation resolutions (semitone scale, velocities coded in 7 bits, etc.) present many limitations for coding musical events. More recent symbolic representation designs such the one used in IRCAM's OpenMusic composition environment [2], provide better abstractions. First, they define data types through an object-oriented formalism, which enables a better data organization, in particular through multiple inheritance of basic types, shown in Figure 3.

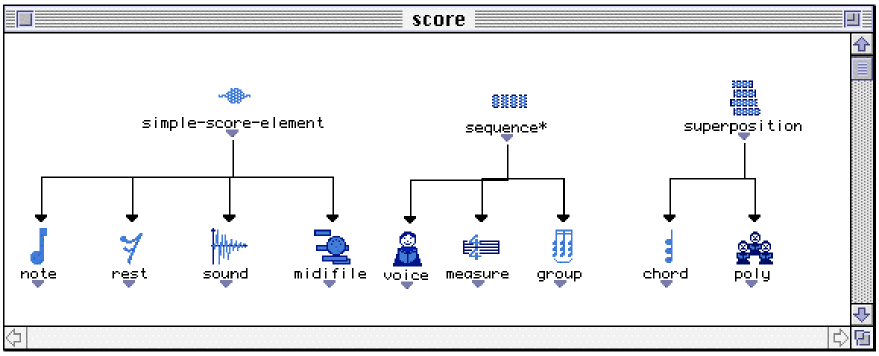

Fig. 3. Basic data types in OpenMusic, from which all music types are derived through multiple inheritance

Second, they enable the formalization of rhythms as hierarchical structures through integer decompositions of a given pulsation, whereas symbolic representations found in sequencers only represent time as a linear and flat axis. As another interesting case, music notation formats as used in score editors, such as NIFF, are actually positioned at a higher level in the RL scale. Let aside all graphic layout information, notation formats obviously combine two different kinds of information□ on one side, discrete musical events (notes) and specifications (e.g. tempo), which are formalized through numerical data and are thus related to the symbolic level; on the other side, textual qualitative specifications which cannot be translated into numerical data, and are hence part of the knowledge level. So music notation formats are actually positioned in the RLs along the symbolic and knowledge levels.

After targeted efforts, in MPEG1, MPEG2 and AAC, to audio compression techniques, recent evolutions of the MPEG standardization process illustrate the need for integrating new representation levels in audiovisual format standards. MPEG4 enables the representation of compound scenes made of various streamed data including compressed audio[8]; it includes, with SASL (Structured Audio Score Language) a format for specifying musical events; moreover, it also includes, in the Advanced AudioBIFS (BInary Format for Scene description), geometrical and perceptual descriptions of 3D audio scenes. Its position in the RL scale thus extends along the physical, signal and symbolic levels.

The integration of data formats associated to the knowledge level is also developing in various standards, including, for example, UNIMARC for digital libraries, MPEG7 [11], which is dedicated to descriptions of audiovisual contents, and MPEG21, which aims at the identification of audiovisual contents and of their right owners.

The SDIF format[3] (Sound Description Interchange Format), developed by several research centers in music technology, is an open format for representing any kinds of audio analysis data (see §3.2). It fills a gap for such representations both within the signal level and at its boundary with the symbolic level.

The respective positions of these standards in the RL grid is illustrated in Figure 4, which exhibits the lack of a performance-oriented control format which would extend MIDI and enable better representations of gestural control information, as part of the physical level, and mappings of these gestures to symbolic information (cf §3.4.3).

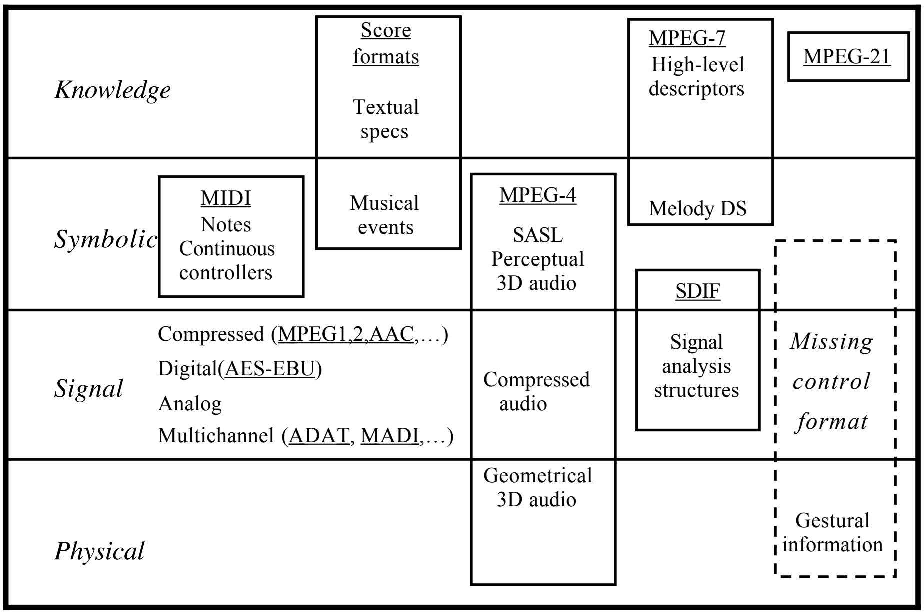

Fig. 4. Mapping of Existing Standards into the RL Scale

3 Analysis of Music Applications through the RL Grid

Since the RL structure is built for providing an extensive view of the various kinds of representations found in music applications, it enables, at least theoretically, to analyze any of them through the grid it provides. This part aims at providing such analysis, through typical examples of music application functions. Therefore, it starts by analyzing usual applications and studying the way they combine the various music

[3] http://www.ircam.fr/equipes/analyse-synthese/sdif/index.html

representations they use. Then, current research issues in music technology are put into the fore as problems of conversion between various levels and through the integration of the physical and knowledge levels.

3.1 Combination of Various RLs in Usual Musical Applications

Usual musical applications, such as sequencers, score editors, audio processing modules, synthesizers, are generally limited to the management of signal and symbolic representations. Moreover, they manage these representations separately, or with limited interactions. This is true with commercial sequencers, which combine audio and MIDI files. Sequencers provide editing and processing functions for each data type, but with interactions between them mainly limited on one side to the audio rendering of MIDI tracks (further referred as the *Synthetic Performance* function) using synthesizers, on the other side to the overall synchronization of all data on a single time reference, which is the basic function of these applications. As a more sophisticated example, the Max application, which results from research performed at IRCAM on real time interaction [18], provides a dataflow processing architecture specifically designed for managing two kinds of musical information□ messages, used in particular for transmitting symbolic information as discrete events, and audio signals. This architecture enables the synchronization of these two data types (Figure 5), through the use of a fixed-size sample block processing architecture for audio signals (typically 64 samples per block), which provides a common clock for all data.

Such an architecture for music information processing, which relies on the synchronization of two clocks, one for audio signals, the other one for discrete and control information, can be actually found in many musical applications. This is the case of DSP plug-ins, i.e. digital audio processing modules designed for commercial sequencer applications.

Functionally, these modules perform the processing of audio signals, through the use of signal processing algorithms, controlled by parameters sampled at a lower frequency rate. This enables to develop sophisticated processing features, which rely not only on signal algorithms, but also on the processing of control parameters. This approach is developed in particular in the GRM Tools processing module suite; each module combines two stages of processing algorithms, the one for audio signal algorithms, the other one for the algorithms parameters, through the systematic use of graphical interfaces [23] and high-level parameter controls. Another conceptual approach of signal processing, developed at IRCAM by Xavier Rodet, relies on an Analysis/□Synthesis architecture, through the use of parametric models. Two distinct model categories are considered: signal models, which describe relevant information through signal processing formalisms, and physical models, which provide acoustic models of sound sources and are hence focused on the causes of sound production, whereas signal formalisms model the effects. Depending on the model formalisms, analytical procedures can be associated to synthesis models. In that case, this enables to derive a specific processing architecture, based on the processing of the parameters, as shown in Figure 6.

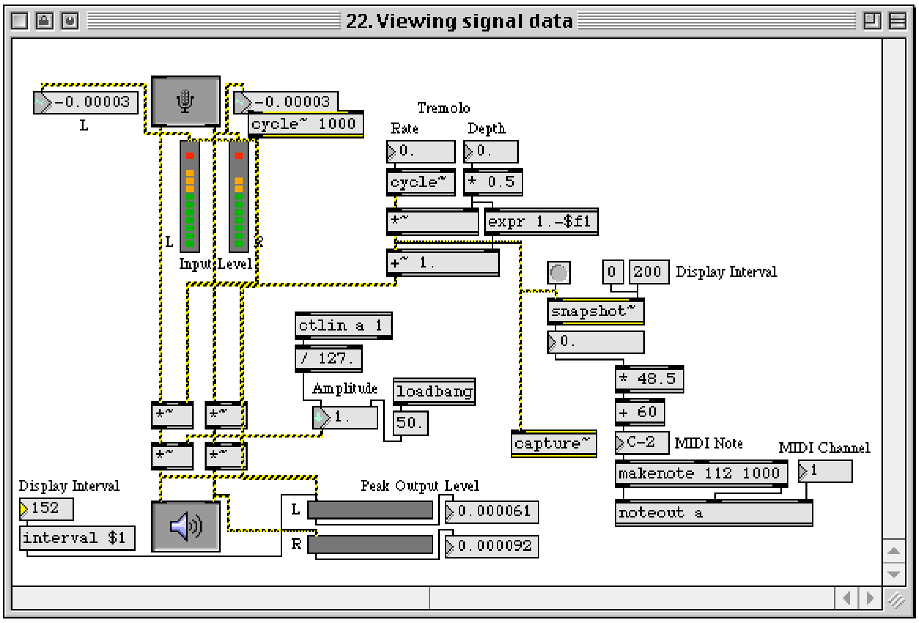

Fig. 5. Example of visual programming patch in Max/MSP. The signal and control processing graphs are respectively represented by dashed and plain lines.

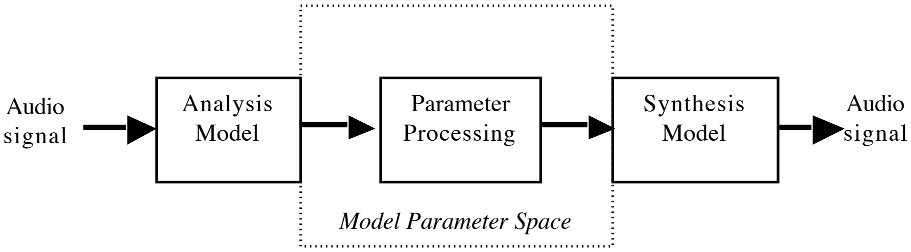

Fig. 6. The Analysis/Synthesis Model Architecture

This architecture is used in particular in the IRCAM Audiosculpt[4] software application, which is based on the Phase Vocoder model (Short Time Fourier Transform). The processing is specified through a graphical interface which displays the model parameters in a time-frequency graphical representation (sonagram), on which the user can draw in order to define editing operations such as time-varying filters (Figure 7).

[4] http://www.ircam.fr/produits/logiciels/audiosculpt-e.html

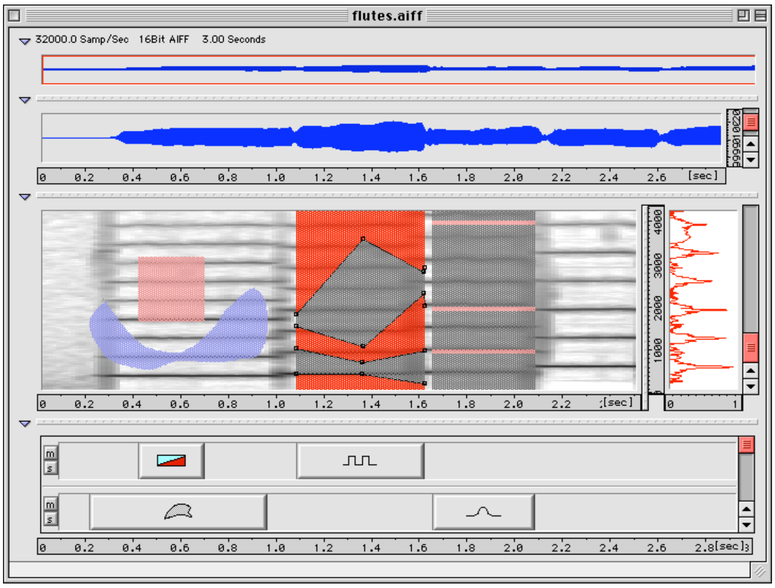
flutes.aiff
32000.0 Samp/Sec 16Bit AIFF 3.00 Seconds
0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 [sec]
1000 2000 3000 4000
0 1
m
s
2.8[sec]

Fig. 7. Time-Frequency Editor of the Audiosculpt 2 Application

In this example, the parameter space (short time Fourier coefficients) actually presents an information quantity of the same magnitude order as the original signal, or even greater. Other models, such as the sinusoidal, or additive model, which decompose an existing signal into a sum of slow-varying sinusoidal functions and a residual signal called noise, enable to have a more quantized parameter space. This space, which includes low-bandwidth amplitude and frequency values of the sinusoidal functions, can be assimilated to the symbolic representation level. In that case, the processing function relies on a dual signal to symbolic (analysis) and symbolic to signal (synthesis) conversion, and the processing is specified in the symbolic space (Figure 8).

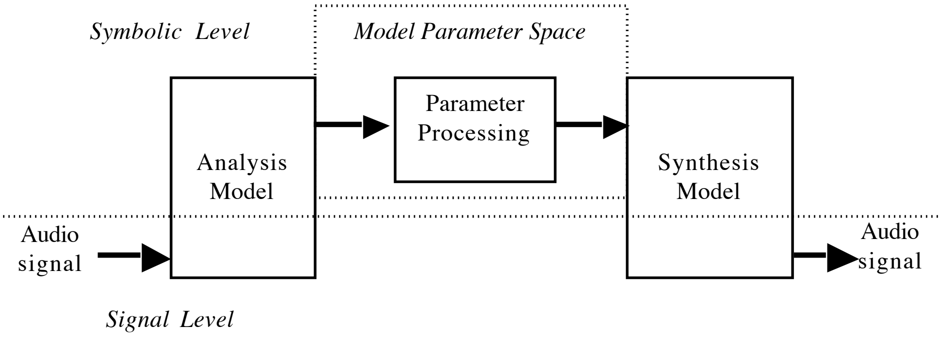

Fig. 8. Audio Processing through Symbolic Information Processing

Such an architecture is widely used in contemporary music productions, which combine recorded audio and instrumental notation materials, through the same symbolic formalism.

3.2 Signal Analysis Issues

The processing scheme of Figure 8 relies on the assumption that an analysis model can be derived from the synthesis model, which is not always the case. Invertible Models, such as the Phase Vocoder, are actually an exception, and deriving analytical procedures from musical signals is currently the subject of many research projects on music technology [6]. Examples of current research issues are listed hereinafter.

3.2.1 Frequency Domain Analysis

Various kinds of information related to the frequency domain can be derived from audio signals: fundamental frequency, spectral peaks, spectral envelopes, etc. In particular, the identification of a limited number of pitch values in a signal mixture provides a digitization from the audio to the symbolic levels. However, in current state of the art in signal analysis, it is still impossible to design robust algorithms capable of analyzing multiple fundamental frequencies present in a polyphonic recording [5].

3.2.2 Automatic Segmentation

The goal of automatic segmentation is to identify time occurrences corresponding to the start and end of musical events. Various time scales can be considered, from the note level (symbolic level) [19] up to the part level in a piece of several minutes (knowledge level) [16]. Some analytical models do not provide only their results as a list of time-stamped events, but also rely on high-level models, such as Hidden Markov Model state sequences: this enables modeling events as compound structures, e.g notes as sequences of various states (Attack, Sustain, Release). However, the unsupervised extraction of musical events from audio signals can be difficult to achieve in some cases. Score alignment algorithms, which perform a synchronization of a reference score expressed in symbolic format with a recorded performance signal, provide more robust results in the general case [15].

3.2.3 Source Separation

As a combination of both former problems, i.e. analysis of superimposed pitches, and temporal analysis, the blind source separation problem aims at separating various sources from a mixture signal. Research in this field [22] is quiet recent and already produces promising results. The goal is to decompose a mixed polyphonic signal into independently varying voices (e.g. instrument groups playing the same voice).

3.2.4 Automatic Identification of Musical Events

The goal of automatic identification is to match information present in the signal with referenced musical events. These events are generally characterized, using machine learning procedures, through the values taken by a vector of numerical descriptors automatically extracted from the signal. An usual form of identification is automatic classification, applied for example to the identification of sound sources (instruments) present in an audio signal, through learned classes characterizing each sound source. Flat and hierarchical automatic classification procedures enable the mapping of identified events to existing taxonomies and can thus be assimilated to automatic conversion functions from the signal level to the knowledge level. New applications, such as digital audio databases, use these automatic indexing features as a computer assistance for classifying sounds in the database [17].

3.2.5 Analysis from Symbolic Representations

Other approaches aim at extracting high-level musical structures from symbolic representations such as MIDI data. Existing research in this field includes automatic meter extraction [20], unsupervised style characterization [1], pattern analysis and matching of melodic, rhythmic and harmonic materials [12]. In the context of music information retrieval, these works, such as the popular query by humming application, generally rely on the analysis of music patterns, between which similarity metrics are applied. One could then wonder where such patterns are positioned in the RL scale, possibly at a missing position between the symbolic and knowledge levels or, like in OpenMusic, as compound structures inherited from basic symbolic objects (notes, chords, event sequences). However, as a specificity of music information, it appears that there is no objective way of producing such patterns from symbolic information; these patterns, such as melodic profiles or surfaces, or rhythmic patterns, only account for a certain aspect of music information, but are not standard elements of the musical syntax. In other words, if compared to language, music includes structural information at the grapheme level (individual notes can be assimilated to individual letters), but not at the lexeme level (there is no structural equivalent to words, as the elementary syntactic and semantic level).

3.3 The Synthetic Performance Problem

We saw in the former examples that a conversion from a low to a higher representation level is done through an analytical procedure, which reduces the information quantity by isolating the appropriate events. Inversely, a conversion between a higher and a lower level will require the generation of additional information, through specific synthesis models. This problem is illustrated through a generic music problem, called here the *Synthetic Performance*application, which aims at simulating the action of a performer and his instrument. In terms of information processing, the performer (plus instrument) function can be summed up as taking symbolic data in input (the score), and producing a variable acoustic pressure at each point of the concert hall space. The computer simulation of this function thus corresponds to a data conversion from the symbolic to the physical levels. A usual way of managing this function in existing applications, generally referred as “MIDI to audio” function, is to assimilate, as shown in Figure 9, the score as a MIDI note sequence, and play it on a synthesizer.

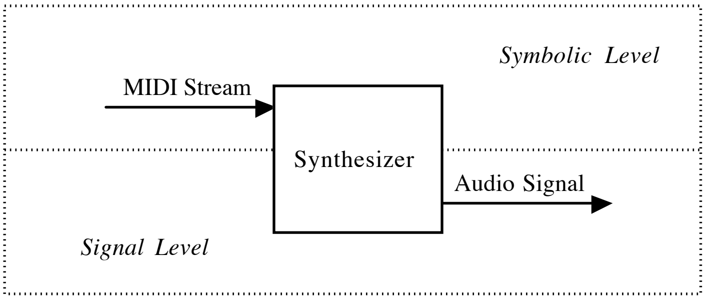

Fig. 9. Basic Performer as MIDI to Audio Function

The synthesizer actually performs the symbolic to audio conversion function by taking symbolic information such as pitch, intensity, duration in input, and generating a signal which simulates a note played by a given instrument. The physical level is thus ignored, as well as most score nuances and indications, and the resulting timbre depends on the selected synthesizer program. A more accurate way of rendering for this Score to MIDI function is to use the General MIDI format, which also transmits, from a standardized instrument set, the instrument to be played to the synthesizer, which then selects the program that best fits the corresponding timbre. A much more extensive view of various processing which should ideally be modeled is displayed in Figure 10.

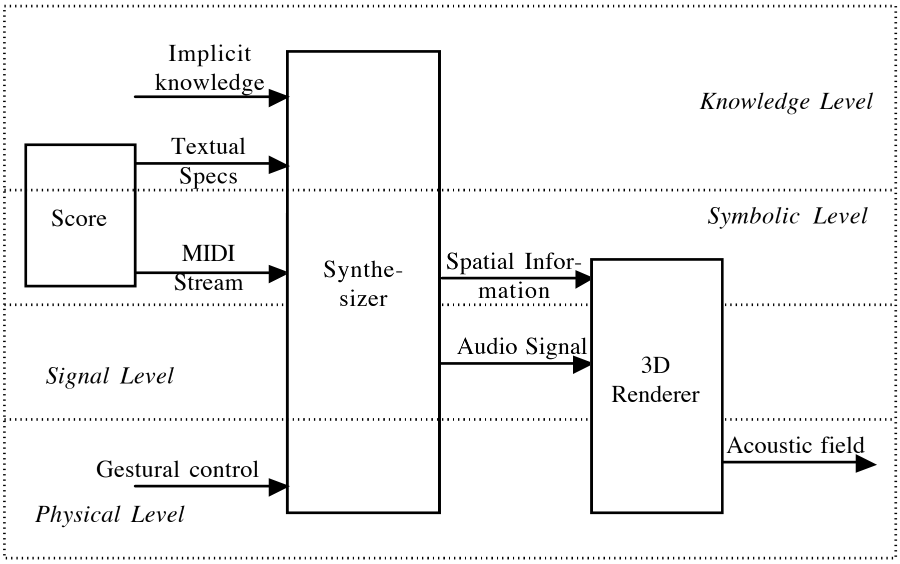

Fig. 10. Ideal Modeling of the Synthetic Performer

It takes into account the two representation levels present in the score, including some textual specifications as part of knowledge level. It also takes into account implicit knowledge, such as cultural references, e.g. existing performances of the same piece. In some specific cases, it also integrates gestural control information relating to the physical level, such as a dynamic model of the performer body, to be coupled with a physical model of the instrument. The physical level is also taken into account through a 3D audio rendering module, which models the instrument radiation, the way this radiation is controlled by the musical text, the simulation of the concert hall room effect, and finally the rendering of all of these spatial effects on the target reproduction system (headphones, stereo, multiple loudspeakers, etc.). This example illustrates the amount of progress to be made in order to better model human performance. Performance modeling is still at its infancy, and, in particular, few studies up to now provide advances in the way score parameters are modulated through the combination of implicit knowledge and explicit textual qualitative specifications.

3.4. Physical Representations

Physical Representations include two kinds of acoustic modeling approaches which have been handled separately up to now□ source models and spatialization models, including room effects. Moreover, it also concerns the representation of gestural information for controlling musical processes (such as synthesis models).

3.4.1. Physical Models of Sound Sources

Many research studies have been dedicated, in the last 20 years, to acoustic models of sound sources, in particular of musical instruments. These models provide powerful sound generation features for composers, such as the Modalys[5] environment, which enables the user to assemble virtual instruments from a set of reference objects (strings, plates, tubes,...) and non-linear interactions (pluck, strike, bow, etc.)[21]. Commercial applications are also developing in the form of real time instrument synthesizers. As compared to signal models, physical models present several advantages□ they produce richer and more realistic sounds; their input parameters correspond to physical variables and are thus more meaningful to the user; the simulation of non-linear coupling functions between the various instrument parts enable the reproduction of the instrument behaviour as a non-linear dynamic system, with possible chaotic output depending on input parameter configurations. However, for this reason, physical models are more difficult to control, and present the same kind of difficulties novice performers experiments when starting to play with instruments with a non-linear behaviour such as winds. Therefore, specific research is done focusing on model inversion, i.e. with the aim of producing a given audio signal through appropriate generation of control parameter values over time. Other research projects are also dedicated to the modeling of instrument radiation, and to the way radiation varies according to played notes.

[5] http://www.ircam.fr/produits/logiciels/modalys-e.html

3.4.2 Spatialization

Spatialization refers to a simulation function which actually includes two different kinds of parameters that had been handled separately in music technology□ localization (sound sources positions), and room effect. Traditionally, these two functions have been simulated respectively using stereo panning (or multi-loudspeaker intensity panning) and artificial reverberation. More recent, physical models enable the complete simulation of the scene from a geometrical description of the room and of the positions of sources and listeners. However, such models, which rely on a convolution of the signal by an impulse response of possibly several seconds, are heavy to compute, and require the impulse response itself to be recomputed as soon as the source or the listener move. Another approach, developed in the IRCAM Spat project [9], is based on a perceptual description of the audio scene. It also combines the localization and room effect simulations, but the room acoustic quality is specified through a set of perceptual parameters. These parameters enable the user to specify the target acoustic quality regardless of the reproduction system, be it headphones (binaural coding), stereophonic (transaural), multi-loudspeakers (intensity panning or Ambisonic), or Wavefield Synthesis. According to the RL grid, the Spat provides a synthesis function from the signal (input sound) and symbolic (target acoustic quality and source positions) levels, to the physical level (reproduction system management). Another, dual approach of spatialization is based on the synthesis of radiation patterns through the control of multi-loudspeaker sources [13]. In the context of live performance and real time processing of acoustic instruments or voice, these sources can be configured in order to approximate the radiation of the processed instrument and thus better fuse with acoustic sounds.

3.4.3 Gestural control

The generalization of electronic instruments can be considered as a continuity of acoustic instrument building. However, electronic instruments bring a rupture for performers in terms of man-machine interaction : acoustic instruments provide a direct energetic coupling between the gestural control and the produced sound, whereas electronic synthesizers get their energy from the electric power and thus bring a decoupling between the control gesture and the synthesized sound [3]. This enables building various kinds of "new instruments" through the combination of any kinds of gesture capture systems and audio synthesizers, and brings a new issue of defining appropriate mappings between gestural control information and the synthesizer control parameters, addressed by several authors [25]. Moreover, the state of the art in live interaction between performing instruments and real time processing systems relies in score following, i.e. real time music shape recognition from a reference score played live by a performer [21]. In order to go beyond the recognition of symbolic information in terms of expressivity, some composers are interested in integrating gestural control information as input signals of musical processes. These evolutions show the need of appropriate representations of musical gestures, as multidimensional signals resulting from captors or image analysis, characterized by a lower sampling rate than audio (typically 1kHz sampling), but also keeping track of cinematic information such as the trajectory, and position, speed and acceleration as functions of time. Beyond MIDI, there is also a need of a new standard for synthesis control, which would account for various representation levels including symbolic information, but also gestural information.

4. Conclusion

This study has identified four main kinds of music representations in existing and potential applications, and has shown that these four types can be organized in levels associated to the various information quantity they convey. These concepts provide a powerful grid for analyzing all kinds of musical applications according to the types of information they manipulate. They also exhibit the lack of a standard syntactic level corresponding to music patterns, even though such patterns are the only way of characterizing structural information between the note level and the high-level form. These concepts are also useful for understanding current issues in music technology research in terms of integration of the physical and knowledge levels and as problems of data conversion between various levels. From bottom up, analysis functions extract relevant information from complex inputs. Inversely, from top down, synthesis functions generate missing information through dedicated models. In both cases, the various illustrations provided from current research show the gap which remains to be filled between functions that could be envisioned for solving basic musical problems and the current limits of our knowledge.

References

1. Assayag, G. Dubnov, S., Universal Prediction Applied to Stylistic Music Generation., In Mathematics and Music, EMS Diderot Forum 1999, Ed. G. Assayag, J.F. Rodrigues, H. Feichtinger. Springer Verlag (2002).
2. Assayag, G., Rueda, C., Laurson, M., Agon, C., Delerue, O.□ Computer-Assisted Composition at IRCAM, From PatchWork to OpenMusic, Computer Music Journal, Volume 23, Number 3, MIT Press (1999) 59-72.
3. Cadoz, C., Continuum énergétique du geste au son, simulation multisensorielle interactive d’objets physiques, In Interfaces homme-machine et creation musicale, Ed H. Vinet and F. Delalande, Hermes Science, Paris (1999), 165-181.
4. Canazza, S., Roda, A., Orio, N., A parametric model of expressiveness in musical performance based on perceptual and acoustic analyses, Proc International Computer Music Conference, ICMA (1999).
5. de Cheveigné, A., Kawahara, H. Multiple period estimation and pitch perception model, IEEE Speech Communication, Vol. 27 (1999), 175-185.
6. Hélie, T., Vergez, C., Lévine, J., Rodet, X., Inversion of a physical model of a trumpet, IEEE CDC : Conference on Decision and Control. Phoenix Arizona (1999).
7. Hiller, L., Music composed with computers, Heinz von Foerster col., J. Wiley & sons, New York (1969).

8. ISO/IEC 14496-1:2000. MPEG-4 Systems standard, 2nd Edition.

9. Jot, J.M., Warusfel, O.□A Real-Time Spatial Sound Processor for Music and Virtual Reality Applications, Proc. International Computer Music Conference, ICMA (1995) 294-295.

10. Mathews, M.V., An acoustic compiler for music and psychoacoustic stimuli, Bell System Technical Journal, 40, (1961), 677-694.

11. Martínez J. M.□MPEG-7 Overview, ISO/IEC JTC1/SC29/WG11 N4980, http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.

12. Meudic, B., St James, E., Automatic Extraction of Approximate Repetitions in Polyphonic Midi Files Based on Perceptive Criteria, Computer Music Modelling and Retrieval, LNCS 2771, Springer-Verlag (2003) *This issue*.

13. Misdariis, N., Nicolas, F., Warusfel, O., Caussé, R.□Radiation control on a multi-loudspeakers device, Proc. International Computer Music Conference, ICMA (2001) 306-309.

14. Orio, N., Lemouton, S., Schwarz, D., Score Following: State of the Art and New Developments, Proc. International Conference on Musical Expression, (NIME-03) (2003), 36-41.

15. Orio N., Schwarz, D., Alignment of Monophonic and Polyphonic Music to a Score », Proc. International Computer Music Conference, ICMA (2001).

16. Peeters, G., La Burthe, A., Rodet, X.,□Toward Automatic Music Audio Summary Generation from Signal Analysis, Proc. International Conference on Music Information Retrieval, IRCAM, Paris (2002).

17. Peeters, G., Rodet, X., Automatically selecting signal descriptors for Sound Classification, Proc. International Computer Music Conference, ICMA (2002).

18. Puckette, M., FTS□ A Real-Time Monitor for Multiprocessor Music Synthesis, Computer Music Journal 15(3), MIT Press (1991) 58-68.

19. Rossignol, S., Rodet, X., Soumagne, J., Colette, J.L., Depalle, P., Feature extraction and temporal segmentation of acoustic signals, Proc. International Computer Music Conference, ICMA (1998).

20. Scheirer, E.,D. Tempo and Beat Analysis of Acoustic Musical Signals J. Acoust. Soc. Am. 103:1, (1998), 588-601.

21. Vergez, C., Bensoam, J., Misdariis, N., Caussé, R.□Modalys : Recent work and new axes of research for Modalys, sound synthesis program based in modal representation□, Proc. 140th Meeting of the Acoustical Society of America (2000).

22. Vincent, E. Févotte, C., Gribonval, R. and al., A tentative typology of audio source separation tasks. Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA 2003), Nara, Japan, (2003).

23. Vinet, H., Delalande, F.: Interfaces homme-machine et création musicale, Hermes Science, Paris (1999).

24. Vinet, H., Herrera, P., Pachet, F.□The CUIDADO Project, Proc. International Conference on Music Information Retrieval, IRCAM, Paris (2002) 197-203.

25. Wanderley, M., Battier, M., Trends in Gestural Control of Music, CDROM book, IRCAM, Paris (2000).