Marshall University
Marshall Digital Scholar

Weisberg Division of Computer Science Faculty Research

Weisberg Division of Computer Science

Summer 7-23-2009

Sound and the City: Multi-Layer Representation and Navigation of Audio Scenarios

Luca A. Ludovico

Davide Andrea Mauro PhD
Marshall University, maurod@marshall.edu

Recommended Citation

Ludovico LA, Mauro DA. "Sound and the City: Multi-layer representation and navigation of audio scenarios." In Proceedings of the 6th Sound and Music Computing Conference. Oporto, Portugal: SMC 2009. 19-24.

SOUND AND THE CITY: MULTI-LAYER REPRESENTATION AND NAVIGATION OF AUDIO SCENARIOS

Luca A. Ludovico, Davide A. Mauro
Laboratorio di Informatica Musicale (LIM)
Dipartimento di Informatica e Comunicazione (DICO)
Università degli Studi di Milano
Via Comelico 39/41 - 20135 Milano (Italy)
{ludovico, mauro}@dico.unimi.it

ABSTRACT

IEEE 1599-2008 is an XML-based standard originally intended for the multi-layer representation of music information. Nevertheless, it is versatile enough to describe also information different from traditional scores written according to the Common Western Notation (CWN) rules. This paper will discuss the application of IEEE 1599-2008 to the audio description of paths and scenarios from the urban life or other landscapes. The standard we adopt allows the multilayer integration of textual, symbolical, structural, graphical, audio and video contents within a unique synchronized environment. Besides, for each kind of media, a number of digital objects is supported. As a consequence, thanks to the features of the format the produced description will be more than a mere audio track, a slideshow made of sonified static images or a movie. Finally, an *ad hoc* evolution of a standard viewer for IEEE 1599 documents will be presented, in order to enjoy the results of our efforts.

1 INTRODUCTION

IEEE 1599-2008 is originally a format to describe single music pieces. For example, an IEEE 1599 document can be related to a pop song, to an operatic aria, or to a movement of a symphony.

Based on XML (eXtensible Markup Language), it follows the guidelines of IEEE P1599, "Recommended Practice Dealing With Applications and Representations of Symbolic Music Information Using the XML Language". This IEEE standard has been sponsored by the Computer Society Standards Activity Board and it was launched by the Technical Committee on Computer Generated Music (IEEE CS TC on CGM) [1].

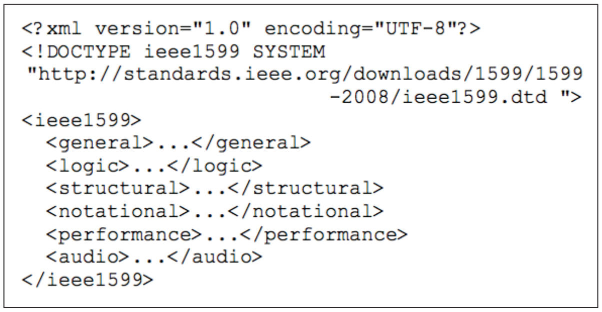
The innovative contribution of the format is providing a comprehensive description of music and music-related materials within a unique framework. In fact, the symbolic score - intended here as a sequence of music symbols - is only one of the many descriptions that can be provided for a piece. For instance, all the graphical and audio instances (scores and performances) available for a given piece are further descriptions; but also text elements (e.g. catalogue metadata, lyrics, etc.), still images (e.g. photos, playbills, etc.), and moving images (e.g. video clips, movies with a soundtrack, etc.) can be related to the piece itself. Please refer to [2] for a complete treatment of the subject. As explained in Section 4, such a rich description allows the design and implementation of advanced browsers.

In this work we are interested in a particular application of IEEE 1599 that goes beyond the original goals of the standard. In fact, instead of applying it to a traditional CWN score, we are going to describe in IEEE 1599 the soundscape of a urban environment using a city map as a score and the different hours of a day to generate different performances.

In the following we will introduce the key features of the standard comparing their traditional meaning in the music field to our new perspective. After, we will describe a generalized viewer for IEEE 1599 format usable both for traditional music pieces and for our goal, namely the audio scenario reproduction. Finally, a case study will be discussed, by using the audio material recorded during 2008 Sound and Music Computing conference held in Genoa (Italy).

Before starting the discussion, a point should be clarified. In our work, a format to encode music information is adapted in order to provide a comprehensive description of a sound environment. This is made possible by the flexibility of the XML encoding we adopt, but it could seem a forcing. We have chosen a format oriented to music since a score is made of symbols corresponding to music events; in our case, the concepts of score and event must be generalized, but the navigation of the audio scenario is similarly driven by events belonging to a predetermined "score". Among many XML-based formats available for music description, IEEE 1599 has proved to be effective, and the reasons are explained in

SMC 2009, July 23-25, Porto, Portugal



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ieee1599 SYSTEM
 "http://standards.ieee.org/downloads/1599/1599
                          -2008/ieee1599.dtd ">
<ieee1599>
  <general>...</general>
  <logic>...</logic>
  <structural>...</structural>
  <notational>...</notational>
  <performance>...</performance>
  <audio>...</audio>
</ieee1599>
```

Figure 1. The XML stub corresponding to the IEEE 1599 multi-layer structure.

Section 2. As shown by the case study, the final result is a multimedia description of city sounds that goes beyond a collection of unrelated materials: the audio environment is depicted at different moments of a typical day; these audio objects are related and synchronized through the occurrence of shared events, which constitute the common score. Furthermore, audio contents are integrated with other multimedia objects both to provide a comprehensive description and to allow innovative ways of browsing.

2 DEFINITION OF MUSIC EVENT

The mentioned comprehensiveness in music description is realized in IEEE 1599 through a multi-layer environment. The XML format provides a set of rules to create strongly structured documents. IEEE 1599 implements this characteristic by arranging music and music-related contents within six layers [3]:

- *General* - music-related metadata, i.e. catalogue information about the piece;
- *Logic* - the logical description of score in terms of symbols;
- *Structural* - identification of music objects and their mutual relationships;
- *Notational* - graphical representations of the score;
- *Performance* - computer-based descriptions and executions of music according to performance languages;
- *Audio* - digital or digitized recordings of the piece.

In IEEE 1599 code, this 6-layers layout corresponds to the one shown in Figure 1, where the root element `ieee1599` presents 6 sub-elements.

The previous list is clearly related to music contents, but in our work layers can be used in a wider context. Before discussing this matter in depth, we have to introduce a key concept of the format, namely the spine.

Since contents are displaced over various levels, what is the device that keeps heterogeneous descriptions together and allows to jump from one description to another? The *Logic* layer contains an *ad hoc* data structure that answers the question. When a user encodes a piece in IEEE 1599 format, he/she must specify a list of music events to be organized in a linear structure called “spine”. Please refer to Figure 2 for a simplified example of spine. Inside this structure, music events are uniquely identified by the `id` attribute, and located in space and time dimensions through `hpos` and `timing` attributes respectively.

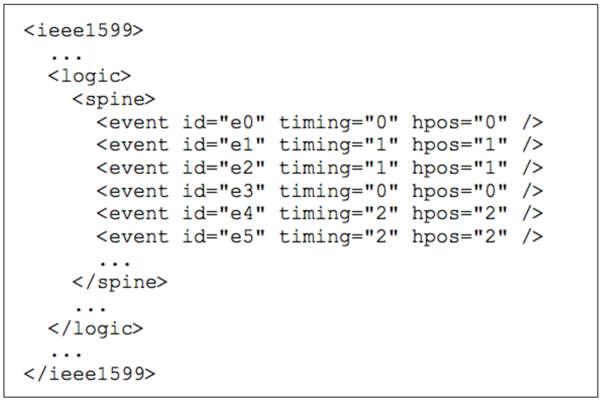
Each event is “spaced” from the previous one in a relative way. In other words, a 0 value means simultaneity in time and vertical overlapping in space, whereas a double value means a double duration of the previous music event with respect to a virtual unit. The measurement units are intentionally unspecified, as the logical values expressed in spine for time and space can correspond to many different absolute values in the digital objects available for the piece.

In the example shown in Figure 2, and regarding it as a music composition, event `e3` forms a chord together with `e2`, belonging either to the same or to another part/voice, as the attributes values of the former are 0s. Similarly, we can affirm that event `e3` should last twice the duration of `e0` (and `e1`), as `e4` occurs after 2 time units whereas `e1` (and `e2`) occurs after only 1 time unit. For further details please refer to the official IEEE draft of the format [4].

In conclusion, the role of the structure known as spine is central for an IEEE 1599 encoding: it provides a complete and sorted list of events which will be described in their heterogeneous meanings and forms inside other layers. Please note that only a correct identification inside spine structure allows an event to be described elsewhere in the document, and this is realized through references from other layers to its unique `id` (see Section 3). Inside the spine structure only the entities of some interest for the encoding have to be identified and sorted, ranging from a very high to a low degree of abstraction.

In the context of music encoding in IEEE 1599, how can be a *music event* defined from a semantic point of view? One of the most relevant aspects of the format, which confers both descriptive power and flexibility, consists in the loose but versatile definition of event. In the music field, which is the typical context where the format is used, a *music event* is a clearly recognizable music entity, characterized by well-defined features, which presents aspects of interest for the author of the encoding. This definition is intentionally vague in order to embrace a wide range of situations. A common case is represented by a score where each note and rest are considered music events. The corresponding spine will list such events by as many XML sub-elements (also referred as *spine events*).

However, the interpretation of the concept of music event can be relaxed. A music event could be the occurrence of a



```
<ieee1599>
  ...
  <logic>
    <spine>
      <event id="e0" timing="0" hpos="0" />
      <event id="e1" timing="1" hpos="1" />
      <event id="e2" timing="1" hpos="1" />
      <event id="e3" timing="0" hpos="0" />
      <event id="e4" timing="2" hpos="2" />
      <event id="e5" timing="2" hpos="2" />
      ...
    </spine>
    ...
  </logic>
  ...
</ieee1599>
```

Figure 2. An example of simplified spine.

new chord or tonal area, in order to describe only the harmonic path of a piece instead of its complete score, made of notes and rests.

Brought to the extreme, the meaning of music event can be extended to comprehend audio events, such as the ringing of church bells or the environmental sounds of a square. Starting from this point of view, our works aims at discovering and exploiting the potentialities of IEEE 1599 format. This a challenging matter as both the format and a number of software tools (e.g. viewers and editors) are already available, but the attempt to apply them to this context is completely original.

3 EVENTS IN A MULTI-LAYER ENVIRONMENT

After giving a correct interpretation to the concept of spine-event, and after the creation of the spine structure, events are ready to be described in the multi-layer environment provided by IEEE 1599. As stated in Section 2, the format includes six layers, which implies 6 families of descriptors for contents.

This section will show that the concept of heterogeneous description is implemented in IEEE 1599 by heterogeneous descriptions of each event contained in spine. While heterogeneity is supported by the whole, inside each layer we find homogeneous contents, namely contents of the same type. The *Audio* layer, for example, can link *n* different performances of the same piece, as well as recordings taken at different times in the same place. In order to obtain a valid IEEE 1599 document, not all the layers must be filled; however their presence provides richness to the description.

In musical terms, the layer-based mechanism allows heterogeneous descriptions of the same piece. For a composition, not only its logical score, but also the corresponding music sheets, performances, etc. can be described. In this context instead, heterogeneity is employed in order to provide a wide range of audio descriptions of the same environment. This concept will become clear in the following.

Now let us focus on the presence and meaning of events inside each layer. The *General* layer contains mainly catalogue metadata that are not referable to single music events (e.g. title, authors, genre, and so on). From the perspective of this paper, the *General* layer could have a poor meaning. Nevertheless, this layer presents a sub-element called `related_files`, a container for 1..*n* specification(s) of external digital objects such as photos, somehow related to the piece but not directly related to the occurrence of music events. For `related_files` sub-element, two attributes are available: `start_event_ref` and `end_event_ref`, containing the identifiers of events listed in spine. These attributes allow to synchronize respectively the appearance and disappearance of static graphical objects with the occurrence of spine events, and they are useful for multimedia presentations. In our case study, we will employ this feature to implement a slideshow of the route, made of images and short text descriptions.

The *Logic* layer, which is the core of the format, faces music description from a symbolic point of view: it contains both the spine, i.e. the main time-space construct aimed at the localization and synchronization of events, and the symbolic score in terms of pitches, durations, etc. The latter aspect is not present in our work; on the contrary the former "logic" description takes a key role for all the other layers, which refer to spine identifiers in order to link heterogeneous descriptions to the same events. Please note that only spine is strictly required by IEEE 1599 format.

Originally, the *Structural* layer has been designed to contain the description of music objects and their causal relationships, from both the compositional and musicological point of view. This layer is aimed at the identification of music objects as aggregations of music events and it defines how music objects can be described as a transformation of previously described objects. Here music events are referred in order to create horizontal (e.g. melodic themes), vertical (e.g. chords), or other aggregations of symbols (e.g. generic segments). In our work, this layer can be used to highlight relationships among events along the route. For example, if two squares are encountered along the way, the *Structural* layer can link the corresponding events. As usual, event localization in time and space is realized through spine references.

For the remaining layers, the meaning of events is more straightforward. The *Notational* layer describes and links the graphical implementations of the logic score, where music events - identified by their spine id - are located on digital objects by absolute space units (e.g. points, pixels, millimeters, etc.). In the case of environmental sounds, the places where they are recorded can be identified over a map. These maps can be the counterpart of the graphical scores as regards our work.

The *Performance* layer is devoted to computer-based performances of a piece, typically in sub-symbolic formats such as Csound, MIDI, and SASL/SAOL. This layer is not used for our goals.

In the *Audio* layer events are described and linked to audio digital objects. Multiple audio tracks and video clips, in a number of different formats, are supported. The device used to map audio events is based on absolute timing values expressed in milliseconds, frames, and so on. Our case study, as discussed in Section 5, includes only three audio tracks, but video clips could be included as well.

Finally, let us concentrate on the cardinalities supported for events layer-by-layer. In the *Logic/Spine* sub-layer, the cardinality is 1 - namely the presence is strictly required - as all the events must be listed in the spine structure. In the *Audio* layer, on the contrary, the cardinality is [0..*n*] as the layer itself can be empty (0 occurrences), it can encode one or more partial tracks where the event is not present (0 occurrences), it can link a complete track without repetitions (1 occurrence), a complete track with repetitions (*n* occurrences), and finally a number of different tracks with or without repetitions (n occurrences). Similarly, the *Notational* layer supports [0..*n*] occurences. In our case, the relevant events listed in spine will be mapped only once for each digital object.

4 BROWSING OF IEEE 1599 DOCUMENTS

In the current section we treat the problem of browsing when many multimedia objects are available, as in the IEEE 1599 environment. Our purpose is presenting a comparison between the standard use with strictly music-related contents and our new application of the format. The interface illustrated in this section represents the evolution of earlier software demos and working applications based on the IEEE 1599 format. It implements the functions and follows the guidelines detailed in [5]. However, till now such an interface has been used only for traditional CWN scores, and in this sense our approach is completely new.

Thanks to the standard, contents can be presented textually, aurally and visually in near real-time to maximize multimedia and multimodal enjoyment of music. In the upper part of Figure 3 an interface for pop songs encoded in IEEE 1599 is proposed. Heterogeneity in music contents is reflected by the layout of controls and views. Players, panels, floating windows or other devices are used to present multimedia contents in a unique framework. Different multimedia types are kept separated by using different controls, whereas objects of the same type are grouped within the same control. For instance, the part of the interface dedicated to audio/video contents contains the playlist of such media objects (dynamically loaded and syncrhonized from the IEEE 1599 file) and the common controls of a media player. Similarly, the panel dedicated to score images contains the list of scores, a control to select the pages of each score (once again dynamically loaded from the IEEE 1599 file) and a number of image-oriented navigation tools.

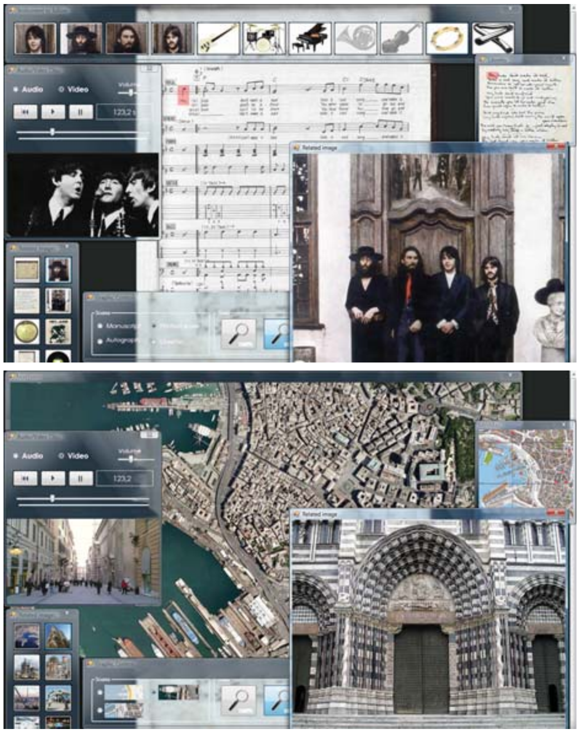


Figure 3. The interface for multi-layer browsing applied to a pop song and to city sounds.

For the goals of this paper the interface has been adapted to the presentation of our material, as illustrated by the lower part of Figure 3. Multimedia and navigation controls can remain unaltered, as the key differences between a music application and this case are not due to a change in media types, rather to a change in the paradigm used to interpret their functions. For instance, now the custom media player loads an audio track of the route and the slider allows to go backward and forward in the audio/video material, whereas the main window (previously used as the “score” panel) contains one of the provided maps of the path itself. The selection tools, that originally have been designed to switch the current score page and music performance, still work in real-time to switch the current map and audio.

Moreover, the interface has been designed to allow the simultaneous enjoyment of all the views involved in the representation of the same piece. Please note that also non-temporized descriptions (e.g. the related files) are accessible. Related files often do not require synchronization, as they are ancillary representations in general not strictly referrable to music events. Usually, in music field this subelement is employed to link on-stage photos, sketches, fashion plates, and so on. Our software application, on the contrary, takes full advantage of related files temporization by showing photos taken during the recording session. The result is a sort of slideshow that enriches the overall description of the route.

5 AN EXAMPLE: ENVIRONMENTAL SOUNDS ALONG A PEDESTRIAN ROUTE

In order to demonstrate the effectiveness of our approach, we have encoded in IEEE 1599 format the results of an experience made during the Sound and Music Computing conference held in Genoa last year (SMC 2008). In that occasion, we recorded the environmental sounds along a short pedestrian path, going from the cathedral (Piazza San Lorenzo, ① in Figure 4) to the harbour (Ponte degli Spinola, ⑥ in Figure 4). This route is about 0.5 km long and it takes about 7 minutes on foot. Sounds were completely acquired three times, namely during three different sessions, trying to respect the same temporization for each capture. We were interested in unveiling the similarities and differences that characterize city life during the phases of a day. To this end, we chose 1am, 9am and 6pm. As a result, we realized that:

- some audio events were quite similar and characteristic for a given place (e.g. the bells of San Lorenzo church), even disregarding time;
- other audio scenarios clearly identified a place or context, but during the day they suffered the consequences of variable human activities (e.g. at the harbour);
- finally, some environmental sounds occurred only in a track (e.g. the transit of an ambulance or the noise of children playing soccer), which adds descriptive richness but decreases the characterizing effect of the audio event over the environment.

Listening to the three mentioned audio tracks was an involving activity indeed, but these digital objects appeared to the listener as something unrelated. In other words, the rationale behind the experiment was clear, but many aspects of interest could have been unveiled only through a synchronization among tracks, the integration with other materials (texts, static images, videos, etc.) and the implementation of *ad hoc* navigation tools to jump from a media to another and to enjoy such a comprehensive description in a unique framework. From this perspective, many similarities emerged with the multi-layer fruition of music provided by IEEE 1599. In that very moment, the idea presented in this paper was born.

Thanks to the features explored in the previous sections, translating our pedestrian route into a city-map based score is easy. Music events identified in spine now become places of interest along the chosen path. The original composition is made of a sequence of music events, as well as a route is made of a sequence of places to visit. Music symbols have a space location over the score, say (x,y) coordinates in pixels, but this information can change from one printed version to another; similarly, the exact localization of places over a map depends on the graphical representation provided by the map itself. In Figure 4 graphical representations have been scaled to make comparisons easier, nonetheless this operation is not required for an IEEE 1599 encoding. However some slight differences are evident between the two upper maps and the lower one. Finally, the original sequence of music events can be translated into different temporized sequences during performances, and this originates a number of audio tracks; similarly, the transit

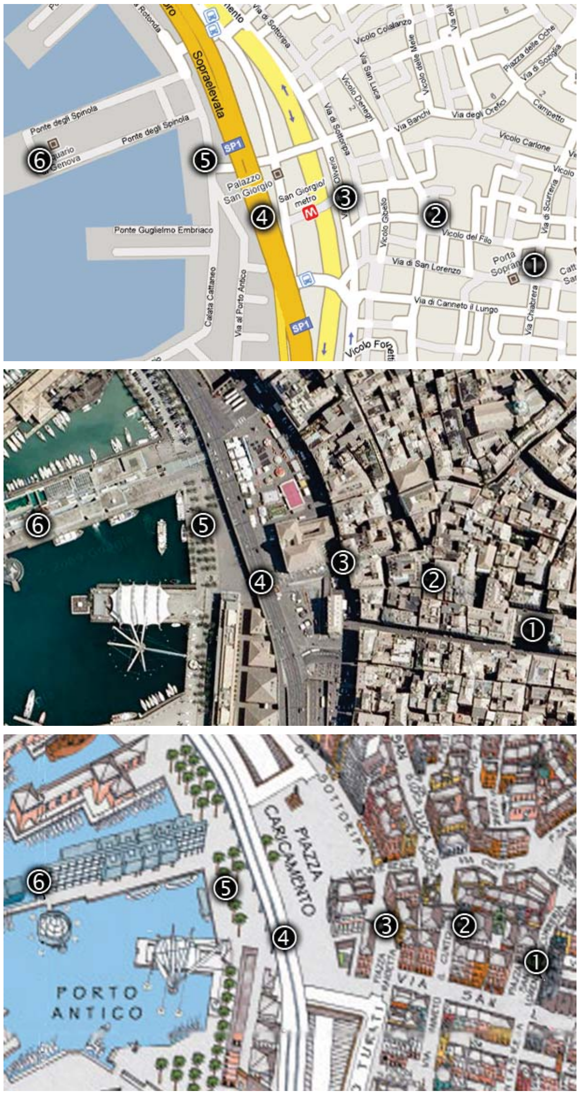


Figure 4. Three representations of the route.

across given map points in general occurs at different moments, and this aspect was captured by our audio recordings.

The points of interest along the route have been marked by numbers. In our path, ① denotes the start point, namely San Lorenzo church, ② identifies the intersection among narrow alleys of the centre, ③ is the crossing with a congestioned avenue, ④ denotes an underpass across the urban elevated motorway, ⑤ identifies the tracking pier for ferries, and ⑥ represents the destination, located at the old dock near the aquarium.

The resulting IEEE 1599 document contains 6 events, listed in the spine structure like in Figure 2. The `hpos` and `timing` attributes have a similar meaning, since the virtual localization in time and space does not refer to a score but to the mentioned pedestrian path. For instance, the relative spacing between each couple of places could be expressed in meters as well as in number of steps. Such events have been mapped within 3 graphical objects and 3 audio objects. For each point of interest, also static images with a text description have been inserted in the *General* layer, in order to generate a slideshow too.

Through this case study, we have proposed only a basic demonstration of the potentialities provided by the format in union with a browsing tool. In broader terms, this experience could be generalized to take into account a number of different scenarios and purposes. For instance, all the main touristic routes of a typical city visit could be represented and proposed in a Web interface to visitors. Another approach consists in encoding sounds not along a continuous path, but statically in a number of places of interest (from a historical, scientific, or other perspective), once again at different moments of either the day or the year. Furthermore, applications to artistic expression and multimedia art installation could emerge.

6 RELATED WORKS

Our work moves from previous experiences such as those cited in [6] and [7]. A number of projects have been carried out about sonification, environmental sounds recording and interaction with city soundscapes. In this sense, we have been explicitly inspired by the “Sons de Barcelona” project by the Grup de Recerca en Tecnologia Musical (MTG) of Universitat Pompeu Fabra - Barcelona. Another source of inspiration is the Freesound project, namely a collaborative database of Creative Commons licensed sounds uploadable and downloadable from the Web.

However, our approach is original as we propose an integrated interface to navigate continuously a map of environmental sounds. Besides, we have explored the use of a new XML-based standard format in order to provide an overall description of city soundscapes. Thanks to the features previously mentioned, IEEE 1599 in our opinion can be efficiently adopted as the format underlying other similar projects.

7 CONCLUSIONS

IEEE 1599 is an XML-based standard originally designed for music pieces. As demonstrated by this paper, the flexibility of the format allows to describe also not-strictly musical contents. We have applied such an encoding to the environmental sounds of a pedestrian route, and developed an application for the visualization and the interaction with multimedia contents. Such an experimental work can be extended in order to provide a virtual visit of an environment driven by a navigable soundtrack.

8 REFERENCES

[1] Baggi, D., “Technical Committee on Computer-Generated Music”, *Computer*, vol. 28, no. 11, 1995, pp. 91-92.

[2] Haus, G. and Longari, M., “A Multi-Layered, Time-Based Music Description Approach Based on XML”, *Computer Music Journal*, vol. 29, no. 1, 2005, pp. 70-85.

[3] Ludovico, L.A., “Key Concepts of the IEEE 1599 Standard”, *Proceedings of the IEEE CS Conference The Use of Symbols To Represent Music And Multimedia Objects*, IEEE CS, Lugano, Switzerland, 2008.

[4] “IEEE Recommended Practice for Defining a Commonly Acceptable Musical Application Using XML”, IEEE, 1599-2008, 2008.

[5] Baratè, A. and Ludovico, L.A., “Advanced interfaces for music enjoyment”, *Proceedings of the working conference on Advanced visual interfaces*, ACM New York, NY, USA, 2008, pp. 421-424.

[6] Gaye, L. and Mazé, R. and Holmquist, L.E., “Sonic City: the urban environment as a musical interface”, *Proceedings of the 2003 conference on New interfaces for musical expression*, National University of Singapore Singapore, Singapore, 2003, pp. 109-115.

[7] Kabisch, E. and Kuester, F. and Penny, S., “Sonic panoramas: experiments with interactive landscape image sonification”, *Proceedings of the 2005 international conference on Augmented tele-existence*, ACM New York, NY, USA, 2005, pp. 156-163.