Towards User-friendly Audio Creation

Cécile Picard
Participating to the numediart Research Program on Digital Art Technologies, Belgium
ccl.picard@gmail.com

Christian Frisson and Jean Vanderdonckt
Université catholique de Louvain (UCL)
Louvain-la-Neuve, Belgium
<first.lastname>@uclouvain.be

Damien Tardieu and Thierry Dutoit
Université de Mons, TCTS Lab
Mons, Belgium
<first.lastname>@umons.ac.be

ABSTRACT

This paper presents a new approach to sound composition for soundtrack composers and sound designers. We propose a tool for usable sound manipulation and composition that targets sound variety and expressive rendering of the composition. We first automatically segment audio recordings into atomic grains which are displayed on our navigation tool according to signal properties. To perform the synthesis, the user selects one recording as model for rhythmic pattern and timbre evolution, and a set of audio grains. Our synthesis system then processes the chosen sound material to create new sound sequences based on onset detection on the recording model and similarity measurements between the model and the selected grains. With our method, we can create a large variety of sound events such as those encountered in virtual environments or other training simulations, but also sound sequences that can be integrated in a music composition. We present a usability-minded interface that allows to manipulate and tune sound sequences in an appropriate way for sound design.

Categories and Subject Descriptors

I.6 [**Information Interfaces and Presentation**]: Sound and Music Computing

General Terms

Algorithms, Design

Keywords

Interactive Sound Composing, Audio Analysis & Synthesis, Content-based Audio Similarity, Multi-fidelity Prototyping

1. INTRODUCTION

Soundtrack composers and sound designers aim at creating auditory experiences [2]. In order to produce soundtracks for movies or video games, Foley artists mainly rely on prerecorded sound material, or record it themselves. While the use of prerecordings is easy to implement, the number of samples in a database is often limited due to memory constraints. Another possibility to generate such sounds is sound synthesis.

A large variety of synthesis methods exist, but each of them is usually more suited for a reduced range of sounds. A very common technique for texture synthesis is the data driven concatenative synthesis, also referred to as mosaicing [11]. Concatenative synthesis approaches aim at generating a meaningful macroscopic waveform structure from a large number of shorter waveforms. They typically use databases of sound snippets, or grains, to create a given target phrase. Unlike granular synthesis where no analysis is performed on the audio units and where the unit size is defined arbitrarily [10], concatenative synthesis selects the audio units according to a set of audio descriptors. Physical modeling can be introduced to further refine granular synthesis [5, 1]. A very important issue for applications of granular synthesis to sound design is the control of the synthesis process. Vocem, introduced by Lopez et al. [7], is one of the first graphical interfaces for real-time granular synthesis, with high-quality audio output and very short latencies. Parameters allow the user to easily control the creation and the distribution of the grains. With MoSevius, Lazier et al. [6] first attempt to apply unit selection to real-time performance-oriented synthesis with direct and intuitive controls based on descriptor values such as energy, spectral flux or spectral centroid, as well as voicing and instrument name. For a more musical context, Misra et al. [8] focus on a single framework that starts with recordings and proposes a flexible environment for sonic sculpting in general. Another class of control methods relies on a wise visualization of the grains database in order to adequately select them. In Catart, Schwarz proposes to display the grains in a two-dimensional space according to descriptor values or output of dimension reduction techniques such as multidimensional scaling analysis or principal component analysis [11].

Following these ideas, we propose an approach that combines hypermedia navigation and a synthesis process into an adequate multimodal user interface for sound composition and design. Our specific contributions are:

- a method for automatic analysis of audio recordings, extraction and classification of meaningful audio grains as new database.
- a technique for automatic synthesis of coherent soundtracks based on the arrangement of audio grains in time.

AM'10, September 15–17, 2010, Piteå, Sweden.

- a usable interface for database manipulation and sound composition.

2. CREATION OF SOUND SEQUENCES

2.1 Method Overview

The proposed method is based on a database of prerecorded sounds. First, the sounds are segmented into small grains. The sounds and the grains are then presented to the user on a graphical interface. Using this interface, the user can select a target sound, that will be used as rhythmic and timbre evolution patterns, and a set of audio grains. Finally, the target and the grains are used to synthesize the new sound using the method described in the following sections. An overview of the synthesis process is given in Figure 1.

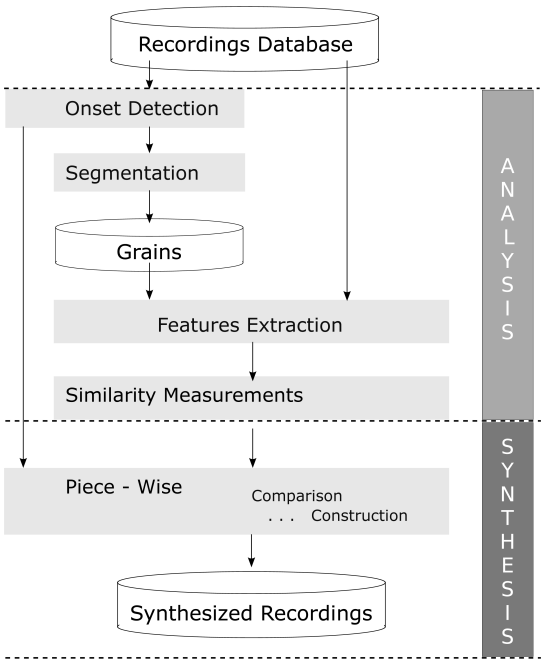

Figure 1: ***Creation of new sound sequences.***

2.2 Extraction of Audio Grains

The first step of the synthesis method consists in the extraction of the audio grains through the segmentation of the database sounds. Segmentation is done using the onset detection method developed in [9]. The method allows the automatic extraction of audio grains whose size is non constant and adapted to the content of the recording through the use of spectral flux measurement. Thus, the technique is able to extract, for instance, discrete drum beatings in a more complex recording of drums. For the sake of clarity, we denote *sound sequences* as long signals characterized by their coherence in the cognitive sense, with a beginning and an end, in contrast to *audio grains*, which are small signals, from 0.01 to 0.1 sec.

2.3 Synthesis method

New sound sequences are built by arranging audio grains in time according to the rhythmic pattern and the timbre evolution of the chosen recording model. We propose to build the synthesized sound sequence piece-wise according to the onset detection of the recording model, by considering the audio grain which is more similar in timbre to the considered excerpt. We use the Mel-Frequency Cepstrum Coefficients (MFCCs) (13 coefficients, 24 filters) to detect timbre similarity, and more precisely we consider the euclidean distance between MFCCs of both the signals. MFCCs are coefficients that collectively express the short-term power spectrum of a sound based on a linear cosine transform of a log power spectrum, on a nonlinear mel scale of frequency. This technique allows us to keep the same timbre evolution of the rhythmic pattern in the synthesized sound sequence. When choosing the audio grain during the synthesis process, we have to consider the case where timbre of the set of audio grains is very different from the rhythmic pattern. In this case, the measure of timbre similarity will always output the same audio grain and the synthesized sound sequence will consequently be built with only one grain. To avoid this bias, we normalize the mean and standard deviation of the MFCC coefficients of both the audio grain set and the segments of the rhythmic pattern.

After having selected the appropriate audio grain, the onset detection of the recording model set the index where the audio grain has to be placed in time. Before concatenating the chosen grain, we apply a tapered window on the grain in order to guarantee the smoothness of the composition.

2.4 Results

The output synthesis depends on the sound material, i.e., the rhythmic pattern and the audio grains. We observe that different tendencies emerge depending on the temporal structure and average timbre of the set of audio grains compared to the rhythmic pattern. If the differences between the recording model and the audio grains are small, the synthesized sound will be similar to the recording model, creating what we call a *coherent variety* within the sound database, otherwise it will create a *disjoined variety*.

Figure 2 shows an example of synthesis using a soundtrack of voice as a rhythmic pattern and audio grains extracting from recording of chimes. When listening to the result, we notice that the timbre evolution and rhythmic scheme of the input pattern are preserved. Figure 2 also specifies the identity of the chosen grains in the concatenation.

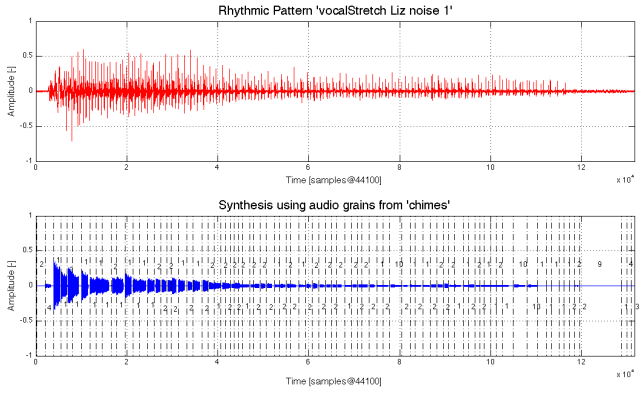

Figure 2: ***Original rhythmic pattern(red) and synthesized sound sequence (blue) based on our synthesis procedure.***

3. DESIGNING A USABLE TOOL

3.1 Framework

The proposed user interface reuses previous achievements gathering within the *MediaCycle* framework for browsing multimedia content by similarity. In this framework, *AudioCycle* is a prototype application dedicated to audio loops libraries developed within the numediart Research Program[1] on Digital Art Technologies since late 2008. The application allows the user to navigate through and listen to audio samples that are visualized and organized according to their similarity in terms of musical properties, such as timbre, harmony, and rhythm. *AudioCycle* thus combines techniques of music-information retrieval, machine learning, information visualization, spatial auditory rendering, and audio time- and pitch-based modification [4].

3.2 Prototyping

As there are very few computerized systems or analog practices that propose a workflow similar to the method we describe here, we had to design a user interface fed by our own creativity. To achieve a certain level of mutual understanding of what we believe to be a suitable design, we produced throughout several brainstormings a storyboard of the expected scenario of usage, as illustrated in Figure 3, and many mockups of the visual user interface, based on previous research [12].
Drawing mockups allowed us at the same time, preventing from diving directly into the implementation of software prototypes. In particular, we avoided a dual-browser solution (one for selecting rhythmic patterns from sound sequences, the second for timbral cues from audio grains) that would have been harder and slower to implement and less straightforward in terms of interaction. Instead, we opted for a single browser displaying temporal and timbral features.

Our user interface is based on a specific scenario of interaction that can be described as follows:

1. browsing, listening to and selecting:
   (a) one sound sequence to be used as rhythmic and timbre evolution patterns for the synthesized sound,
   (b) several audio grains to be used as timbral material,
2. easily composing a new sound sequence based on the recording model, creating *coherent* or *disjoined* variety according to the timbral content of the chosen audio grains compared to the one of the recording model,
3. listening to the new sound sequence, optionally saving it (and thus making it appear on the browser),
4. renewing the aforementioned cycle (steps 1. and 2.), by either choosing another recording model or different grains, or starting again with no audio content set.

3.3 Towards a usable tool

Since the first stages of the MediaCycle framework, simple geometrical shapes have deliberately been used to visually represent the elements of the media database in its browser.

[1] `http://www.numediart.org`

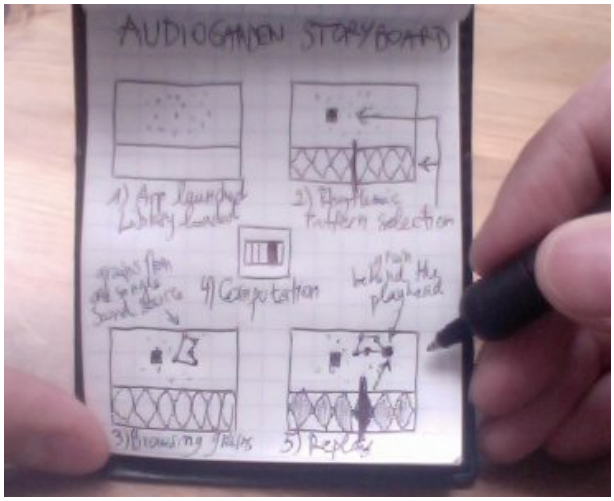

Figure 3: ***Storyboard of an expected scenario of usage quickly drawn on a small notebook.***

Some variations and artifacts are used (with information visualization techniques and visual variables assignations [13]) so as to underline: particular relations between elements (node-links), classes/clusters of elements (colors), type of media (shapes), saliency of some elements at a given user interaction step (distortion, fisheye), and so on... The low visual complexity allows us therefore quite conveniently to narrow down the boundaries between paper mockups, vectorial renderings, and the actual software prototypes, so that it is easier to keep users in the loop while discussing and assessing the viability of the proposed solutions during these various stages of multi-fidelity prototyping [3].

The current version of the software prototype presents a two-pane interface that gathers a visual display (right pane), and control parameters for browsing audio libraries and synthesis computation (left pane), see Figure 4. The right pane is divided in two parts as follows:

- at the top, a visual display of the sound sequences and extracted grains present in the loaded audio database; placement in the 2D space is based on timbral characteristics and recording duration;
- at the bottom, a visual display of the sound being synthesized, starting from the waveform of the chosen recording model, then moving towards the synthesized sound sequence as audio grains are added to the synthesis process.

In the left pane, parameter controls allow loading and saving audio files/libraries, but also setting, starting and stopping an OSC client/server for interaction with gestural controllers. Browser navigation can be set through specific similarity parameters, such as timbre and/or rhythm. Synthesis settings are proposed for sound composition: the user can select different modes of arranging audio grains in time and can add randomness in the audio grains sequence, taking thus some distance from the recording model regarding the rhythmic pattern and the timbre evolution.

We provide with a video[2] that further illustrates our method.

[2] Online video: `http://www.dailymotion.com/video/xe8ao0_numediart-10-2-audiogarden_tech`

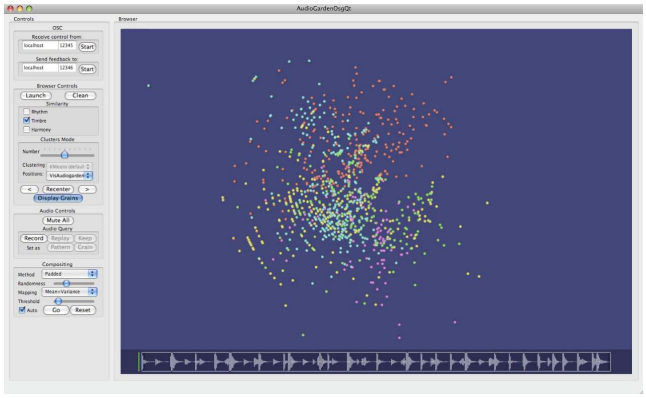

Figure 4: Screenshot of an early prototype of the user interface, featuring a two-pane view: browser controls in the left pane, a composite visual display in the right pane whith a visual display of the loaded audio database (top), and a visual display of the sound being synthesized (bottom).

We used samples from the One Laptop Per Child (OLPC) Free Sound Samples Library [3]. The video shows an example of synthesis using a recording of anklung instrument as a model for rhythmic pattern and timbre evolution, and audio grains extracted from a soundtrack of music box.

We have to ensure users have a pleasurable experience when using the tool. Once we reach a mature version of our prototype, we plan to undertake qualitative usability tests, with tasks such as trying to recreate at best a target sound sequence that the user has just listened to. The results would assess the efficiency and rapidity of the tool.

4. DISCUSSION AND CONCLUSION

We propose a new approach to soundtrack creation and sound design, with applications to virtual environments, by offering a tool that allows sound navigation and manipulation, and flexible composition. Our method processes a database of prerecordings for segmentation and semi-automatically composes the obtained audio grains according to a given recording model, following rhythmic and timbre evolution coherence. Our technique can create a large variety of sound sequences according to the chosen sound material. This novel algorithm for sound sequence creation has been integrated in a first prototype tool, meant to be pleasurable and efficient, that allows to browse sounds, display them accordingly in a signal-based classification (in terms of timbral and/or rhythmic features), and select one sound sequence and several grains so as to create new sound sequences.

For future work, we aim at gathering physically based synthesis methods and handling of prerecordings in the same interface. We will propose to address physically based synthesis through modal analysis, which consists in modeling a vibrating object by a bank of damped harmonic oscillators excited by an external stimulus. Modal decomposition allows efficient runtime and control for vibrational response. Modal sounds can be considered themselves as audio grains, and they can be consistently introduced into the database classification. Physically based sounds may allow the user to *design* from a shape, as for example with a CAD model, to its sound and incorporate it into a sound composition. The user interface still needs to be refined: visualization techniques need to be carefully chosen in order to make visual points of interest more salient, and the control of the navigation in the database and of the sound compositing would benefit from a dedicated gestural interface.

[3]One Laptop Per Child Free Sound Samples Library: `http://wiki.laptop.org/go/Sound_samples`

5. ACKNOWLEDGMENTS

This work has been mostly financed by `numediart`, a long-term research program centered on Digital Media Arts, funded by Région Wallonne, Belgium (grant N°716631). C. Picard obtained a Short-Term Scientific Mission (STSM) funding from the COST Action Sonic Interaction Design (SID). We want to acknowledge OLPC for providing with the Free Sound Samples Library under Creative Commons license.

6. REFERENCES

[1] P. R. Cook. Toward physically-informed parametric synthesis of sound effects. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1999.

[2] C. Cox and D. Warner, editors. *Audio Culture: Readings in Modern Music*. Continuum International Publishing Group, 2004.

[3] A. Coyette, S. Kieffer, and J. Vanderdonckt. Multi-fidelity prototyping of user interfaces. In *Proceedings of INTERACT*, 2007.

[4] S. Dupont, C. Frisson, X. Siebert, and D. Tardieu. Browsing sound and music libraries by similarity. In *128th AES Convention*, London, UK, 2010.

[5] D. Keller and B. Truax. Ecologically-based granular synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, Ann Arbor, USA, 1998.

[6] A. Lazier and P. Cook. MOSIEVIUS: Feature driven interactive audio mosaicing. In *Proceedings of the International Conference on Digital Audio Effects*, London, UK, 2003.

[7] D. Lopez, F. Marti, and E. Resina. Vocem: An application for real-time granular synthesis. In *Proceedings of the Digital Audio Effects(DAFx)*, 1998.

[8] A. Misra, P. R. Cook, and G. Wang. Musical tapestries: Re-composing natural sounds. In *Proceedings of International Compupter Music Conference (ICMC)*, New Orleans, USA, 2006.

[9] C. Picard, N. Tsingos, and F. Faure. Retargetting example sounds to interactive physics-driven animations. In *Proceedings of the AES 35th International Conference on Audio for Games*, 2009.

[10] C. Roads. Introduction to granular synthesis. *Computer Music J.*, 12(2), 1988.

[11] D. Schwarz. Concatenative sound synthesis: The early years. *Journal of New Music Research*, 35(1):3–22, 2006.

[12] C. Snyder. *Paper Prototyping: The Fast and Easy Way to Define and Refine User Interfaces*. Morgan Kaufmann, 2003.

[13] C. Ware. *Visual Thinking: for Design*. Interactive Technologies. Morgan Kaufmann, 2008.