The Sound Space as Musical Instrument: Playing Corpus-Based Concatenative Synthesis

Diemo Schwarz
UMR STMS
Ircam–CNRS–UPMC
Paris, France
schwarz@ircam.fr

ABSTRACT

Corpus-based concatenative synthesis is based on descriptor analysis of any number of existing or live-recorded sounds, and synthesis by selection of sound segments from the database matching given sound characteristics. It is well described in the literature, but has been rarely examined for its capacity as a new interface for musical expression. The outcome of such an examination is that the actual instrument is the space of sound characteristics, through which the performer navigates with gestures captured by various input devices. We will take a look at different types of interaction modes and controllers (positional, inertial, audio) and the gestures they afford, and provide a critical assessment of their musical and expressive capabilities, based on several years of musical experience, performing with the CataRT system for real-time CBCS.

Keywords

CataRT, corpus-based concatenative synthesis, gesture

1. INTRODUCTION

Corpus-based concatenative synthesis (CBCS) is a recent method for sound synthesis [5, 6], that has been implemented since 2005 in an interactive sound synthesis system named CataRT, used in sound design, composition, and installation contexts, and used by the author (see `http://music.concatenative.net`) and other musicians for live music performances [7].

While the technological and scientific bases of CBCS have been well described and theorised, its use as a new interface for musical expression has not been treated specifically. Yet, it introduces an important and novel concept that is the essence of the interface: the space of sound characteristics with which the player interacts by navigating through it, using gestural controllers. Therefore, this article will try to take a first step towards formalising the experience of this use of CBCS as a musical instrument made by the author mainly in a setting of improvisation with other musicians.

We will start by giving a short general introduction to the principle of CBCS and mention some related approaches, before investigating the central notion of this article, the sound space as an interface to CBCS in section 2, and various gestural controller devices that allow to interact with it, each with certain advantages and disadvantages, in section 3. After that, we turn to examining the types of gestures afforded by these interfaces, and the different trigger modes in section 4, before investigating how to construct the sound space, and optimise its representation in section 5. Finally we provide a critical assessment of the different variants of the instrument in section 6 followed by general conclusions and avenues for future work in section 7.

In order to better convey the interactive and gestural aspects of the interfaces, a companion web-page at `http://imtr.ircam.fr/imtr/CataRT_Instrument` gives video and audio examples for each of the controllers.

1.1 Principle of CBCS

Corpus-based concatenative synthesis systems build up a database of prerecorded or live-recorded sound by segmenting it into *units*, usually of the size of a note, grain, phoneme, or beat, and analysing them for a number of sound descriptors, which describe their sonic characteristics. These descriptors are typically pitch, loudness, brilliance, noisiness, roughness, spectral shape, or meta-data, like instrument class, phoneme label, that are attributed to the units, and also include the segmentation information of the units. These sound units are then stored in a database (the *corpus*). For synthesis, units are selected from the database that are closest to given *target* values for some of the descriptors, usually in the sense of a weighted Euclidean distance. Non real-time CBCS [6] can also make use of a *unit selection* algorithm based on dynamic programming or constraint solving [11] that finds the sequence of units that match best a given sound or phrase to be synthesised (the target). The selected units are then concatenated and played, after possibly some transformations.

1.2 Motivations for CBCS

Ever larger sound databases exist on all of our harddisks and are waiting to be exploited for synthesis, which is ever less feasible to do completely manually. Therefore, the help of automated sound description allows to access and exploit a mass of sounds efficiently and interactively, unlike traditional query-oriented sound databases [8].

As with each new synthesis method, CBCS gives rise to new sonorities and new methods to organise and access them, and thus expands the limits of sound art.

Last, using concatenative synthesis, as opposed to pure synthesis from a signal or physical model, allows a sound composer to exploit the richness of detail of recorded sound while retaining efficient control of the acoustic result by using perceptually and musically meaningful descriptors to specify the desired target sound features.

1.3 The CataRT System

The modular CataRT free software system [7] for Max/MSP with the FTM&Co extensions, available at `imtr.`

NIME'12, May 21 – 23, 2012, University of Michigan, Ann Arbor.

ircam.fr and ftm.ircam.fr, realises CBCS in real-time. It analyses any number of sound files or live audio, segmenting by fixed size, by pitch or spectral change, by attack detection, or by importing markers. For synthesis, the target of the selection is most intuitively controlled in a 2D representation of the corpus, where each unit is a point that takes up a place according to its sonic character. Access via more dimensions is also possible.

CataRT must be seen as a whole family of possible instruments, around the core concept of the sound space. The actual instrument and the interaction it affords is determined by the controller that steers the navigation, and the choice of gestural interaction and trigger modes (section 4).

1.4 Related Work

CBCS and related approaches have been developed independently in a number of projects, summarised in a survey [5] that is constantly kept up-to-date at http://imtr.ircam.fr/imtr/Corpus-Based_Sound_Synthesis_Survey. The most prominent related approach is that of *audio mosaicing* [11], which is a special case of CBCS, when the selection target is derived from another audio signal. Most often, the descriptors used in that case are a representation of the audio spectrum in a low number of frequency bands, resulting in a lookup of sounds from the corpus by timbral match. CBCS can also be seen as a content-based extension to granular synthesis providing direct access to grains with specific sound characteristics in real-time, thus surpassing its limited selection possibilities, where the only control is position in one single sound file.

2. THE SOUND SPACE AS INTERFACE

By now it should be clear that the central notion of CBCS as an instrument resides in the representation of the corpus as an abstract space of sound characteristics, spanned up by n (24 for CataRT) audio and meta-data descriptors as its axis. In this space, similar sounding units are close together, and large distances mean that the sounds are very different (within the limits of what can be captured by the descriptors). Note that this concept is similar but not equivalent to that of the *timbre space* put forward by Wessel [10], since timbre is defined as those characteristics that serve to distinguish one sound from another, that remain after removing differences in loudness and pitch. Our sound space explicitly includes those differences that are very important to musical expression.

The sound space is populated by *units* of sound, placed at the coordinates given by their audio descriptors or class indices. In order to make this high-dimensional representation into an interface, the visual representation is a projection into a lower-dimensional space—two or three dimensions are best adapted to the commonly available display and control devices (see figure 1a for an example).

While already the indexing of the corpus by only a few high-level perceptual descriptors allows a musical and close-to-symbolic access to sound, the further projection of the descriptor space to a low-dimensional navigation space makes the interaction with the sound corpus very straightforward and intuitive.

Here, playing means navigating through the space, (whereas in audio mosaicing, playing means querying the space by audio examples), and triggering the segments closest to the current position according to one of several possible trigger modes that determine the gestural affordances of the interface, detailed in section 4.

3. CONTROLLERS

The actual instrument is determined by the controller that steers the navigation, which fall into the groups of positional control (3.1), and control by the analysis of audio input (3.2). Additionally, granular playback and transformation parameters are conveniently controlled by faders.

This section will describe the controllers from a technical point of view. The following section 4 will link their input to the gestures they afford, while section 6 will compare and discuss them.

3.1 Positional Control

This is the straightforward control method for CataRT, mapping a 2D or 3D controller to the target position.

For composition and micro-montage, even standard pointing devices such as the mouse or trackpad can do, as seen in video 1.1, however they lack possibilities of dynamic play.

3.1.1 XY–Controllers and Surfaces

The most intuitive access to CataRT is provided by XY controllers, such as MIDI control pads (like the KAOSS pad), MIDI joystick controllers, etc., for giving the target position in 2D. Better still are pressure-sensitive XY-controllers such as a graphics tablet (Wacom), or the STC-1000, to control also dynamics.

The graphics tablet allows a very precise positioning of the target via a pen, and can register (light) pressure. Additionally, the tilt of the pen can be used to control two more parameters. See video 3.1 and 3.2.

The Mercurial Innovations Group STC-1000 MIDI controller, now no longer produced, has a Tactex surface with a grid of 3x3 pressure sensors. It can send the raw 1024 bit pressure data that allows to recalculate the position with higher precision as the standard output via the center of gravity of the pressure values. One can also determine the size of the touched area, given by the standard deviation. This allows to control one additional parameter by the distance between two fingers. See video 4.1.

3.1.2 Multi-Touch Surfaces

Multi-touch controllers, like the Jazzmutant Lemur or the touch-screen prototype Stantum SMK, and the pressure sensitive Continuum Fingerboard or the planned Linnstrument, are the dream interface for CataRT, providing polyphonic access to a sound space. The Stantum SMK has been used for a one-off performance [1], see videos 5.1–5.3.

3.1.3 Motion Capture

Motion capture systems, either by cameras and markers, or the Kinect, offer a full-body access to a sound corpus mapped into physical 3D space. These interfaces have not yet been used for music performance with CataRT, but are beginning to be used in installation settings.

3.1.4 Accelerometers

Accelerometer equipped devices such as Wiimotes, smartphones, or tablets can be used to navigate the sound space by tilting and shaking (see video 6.1). In the mapping used in CataRT, this falls within positional control, since the accelerometer readout of the device is mapped to 2D position: held flat, the target is in the middle of the navigation space, tilting moves it towards the gradient. However, the mass and inertia of the device add a physical component to the interaction, making the gestural affordance of the controller very different from above pointing devices.

3.2 Audio Control

3.2.1 Piezo Microphones

Piezo pickups on various surfaces allow to hit, scratch, and strum the corpus of sound, exploiting all its nuances according to the gestural interaction, the sound of which is anal-

ysed and mapped to the 2D navigation space of CataRT see videos 8.1 and 8.2: The approach here uses an attack detector (`bonk~`) that also outputs the spectrum of the attack audio frame in 11 frequency bands. Total energy and centroid of this spectrum is mapped to the x and y target position in the 2D interface to select the units to play from the corpus. This means, for instance, dull, soft hitting plays in the lower-left corner, while sharp, hard hitting plays more in the upper right corner. The attack detection is not 100% accurate, but since the signal from the piezos is mixed to the audio played by CataRT, the musical interaction still works.

3.2.2 Audio Analysis

The descriptor-analysis of incoming audio can serve to control the selection of similar-sounding units [9]. A mapping stage can automatically adapt the range of the input descriptor to the corpus, in order to exploit it fully, forsaking precise lookup of pitch, for instance, in favour of gestural analogies. This possibility of mapping makes for a significant difference with the control of CBCS by audio spectrum analysis as in audio mosaicing, described in section 1.4.

4. GESTURAL INTERACTION

The controller devices explained in the previous section provide the target position in the 2D navigation space and possibly one or more dynamic parameters such as pressure. How and when the units are actually played is subject to the chosen trigger mode that, together with the control device, finally determines the gestural interaction. We will analyse the gestures within the framework of Cadoz [2], that distinguishes the three types of *excitation*, *selection*, and *modification* gesture. There are two groups of interaction that give rise to two different playing styles as follows.

Excitation by Selection In the first group of gestural interaction, the specification of the target position can at the same time trigger the playback of the unit, i.e. the navigation in the sound space constitutes a combined selection and an excitation gesture.

In CataRT's *fence* trigger mode, passing over a unit triggers it. Swiping gestures can trigger arpeggios of units, the rate and density of which depend on the speed of the swiping gesture. The *grab* trigger modes "grabs" the currently closest unit while a button is pressed (a modification gesture) and triggers it repeatedly with a rate proportional to the speed of movement of the target position. Additionally, because the position has no selection function during grab, we can map it, for instance, to spatialisation.

With controllers that afford pressure sensitivity, very dynamic, yet precisely controlled, sonic gestures can be produced by effecting different trajectories, speeds and pressure profiles on XY controllers. With accelerometers, the speed of tilting, or the vigorousness of shaking the controller is proportional to the density of the produced sound, while the precision of the selection is much lower.

Regular Excitation This other group of gestural interaction separates selection from excitation, giving rise to continuous rhythms or textures. Here, either a metronome triggers units regularly (*beat* and *quant* trigger modes), or one or more units are looped (*chain* or *continue* mode).

The former can produce dense sonic textures, the timbre of which can be precisely chosen by the target position. Their variability or steadiness can be influenced by the selection radius, and the random ranges of the granular playback parameters. The latter recalls parts of the corpus sound files, restarting by giving a new target position.

Here, the navigational gestures are solely selection gestures, while no excitation gestures are needed, since the system plays continuously. However, the trigger rate and the granular playback parameters are controlled by modification gestures on faders.

5. BUILDING THE SPACE

We now turn our attention to three aspects of setting up the sound space for use as a musical instrument: How to fill it with sounds, how to organise them by classes and voices, and how to optimise the 2D/3D layout for interaction.

5.1 Filling the Corpus

The corpus can be built either from any number of existing sound files, or can be recorded live during the performance from a musician or environmental sources, giving rise to live CBCS [3], very appropriate for improvised performances:

Here, starting from an empty corpus, CataRT builds up the database of the sound played live by segmenting the instrument sound into notes and short phrases. The laptop performer then re-combines the sound events into new harmonic, melodic and timbral structures.

5.2 Organising the Space

In order to create, and precisely and expressively control variety in the played sounds, the sound space can be organised in three aspects: First, any number of separate corpora can be loaded or recorded in CataRT and assigned to one of the synthesis channels. This provides completely separate sub-instruments, each with its own sound space, between which the performer can switch by modification gestures.

Then, within one corpus, sound files or individual units can be grouped into *Sound Sets*, which can be selectively enabled or muted, i.e. excluded from selection. When this choice is accessed by a modification gesture, e.g. on buttons, quick changes in the played sounds can be effected, similar to a percussionist changing instrument or mallets.

Last, each synthesis channel in CataRT can generate several independent voices, with separate trigger mode and granular transformation parameters, allowing to simultaneously play from a corpus with a metronome and with dynamic gestures, for instance, or with different filter settings. This is comparable to the clean and crunch channels on guitar amplifiers, or a multi setup on a synthesiser keyboard.

5.3 Optimising the Navigation Space

While a direct projection of the high-dimensional descriptor space to the low-dimensional navigation space has the advantage of conserving the musically meaningful descriptors as axes (e.g. linear note pitch to the right, rising spectral centroid upwards), we can see in figure 1a that sometimes the navigation space is not optimally exploited, since some regions of it stay empty, while other regions contain a high density of units, that are hard to access individually. Especially for the XY controller in a multi-touch setting, a lot of the (expensive and always too small) interaction surface can remain unexploited. Therefore, we apply a distribution algorithm [4] that spreads the points out using iterative Delaunay triangulation and a mass–spring model, the results of which are illustrated in figure 1b.

This last method starts from a 2D projection and optimises it. Alternatively, we can already integrate the high-dimensional descriptor similarity in the projection to 2D, by using the hybrid multi-dimensional scaling algorithm [8]. mass–spring model.

6. DISCUSSION

In lieu of a formal evaluation, this section will try to provide a critical assessment of using the sound space as an interface for musical expression, and notably the CataRT sys-

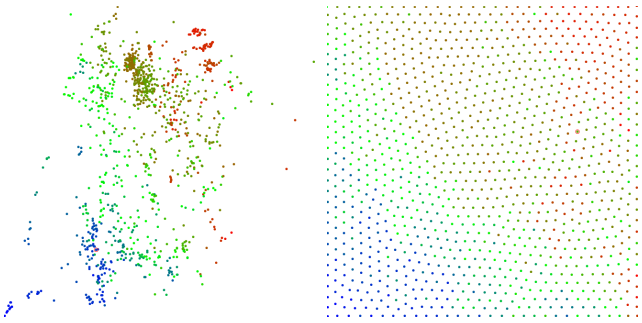

Figure 1: (a) Corpus plotted by Spectral Centroid (x), Periodicity (y), NoteNumber (colour), (b) *Unispring* **distribution of the corpus.**

tem with its various controllers, based on the performance experience of the author.

First, regarding musical expressiveness, the dynamic gestural trigger modes, combined with varied corpora, allow a great range of dynamics and densities to be produced. Throwing around clouds of grains, imprinting them with timbral evolutions, is one of the strong points of CataRT.

The continuous playing modes are also a staple for multi-layered musical scenes, where complex textural soundscapes can be built up out of virtually any source material, and finely yet dynamically controlled.

A weak point is rhythmic play, a steady beat can be generated, but is nearly impossible to vary enough to be musically interesting; neither is tight percussive play, because of two sources of imprecision: The segmentation of the units, especially in live recording, might not be close enough to an actual attack of the segment, and the controllers all introduce a latency too noticeable for percussion.

Turning to the controllers and their comparative advantages and disadvantages, a first manifest difference in attitude can be seen in the play using a graphics tablet and a pressure-sensitive xy-pad: The use of a pen for the graphics tablet will allways lead to writing-like gestures, minute movements with high precision will have large effects, and will be played sitting down. This stands against the physical force of pressure exerted on controllers like the STC-1000, being played standing, that allows to express musical dynamics with implication of the full body. Even the (non pressure sensitive) multi-touch prototypes could not render this physical implication, although they allow virtuoso multi-finger gestures and bi-manual play.

The piezo audio control permits very expressive play, sometimes creating gestural analogies, e.g. to violin playing, but much less precision, since the gestures are funneled through an attack detector.

A more general questioning of the concept of the sound space as interface is the antagonism of large variety vs. fine nuances, that need to be accomodated by the interface. Indeed, the small physical size of the STC-1000 does sometimes not provide a fine-enough resolution to precisely exploit fine nuances. Here, prepared sound sets and zooming could help, but finally, maybe less is more: smaller, more homogeneous corpora could invite to play with the minute details of the sound space.

This reflection also leads to the only seemingly problematic point that the interface relies on visual feedback to support the navigation in the sound space, and that this feedback is on a computer screen, separate from the gestural controller (except for the multi-touch screen from Stantum). Indeed, for fixed corpora, this can be easily circumvented by memorising the layout and practising with the corpora for a piece, as has been shown in the author's interpretation of the piece *Boucle #1* by composer Emmanuelle Gibello, where the computer screen is hidden.

Last, performances using live corpus-based concatenative synthesis create a very special improvisational relationship, with very strong coupling between the acoustic and the electronic performer [3]. Here, the sound space becomes a shared instrument, nourished by the former, and consumed by the latter, creating a symbiotic relationship between the two musicians.

7. CONCLUSION AND FUTURE WORK

The sound space is a powerful metaphor for organising and accessing an enourmous wealth of sounds, while still retaining precise control about timbre and dynamics of sound production. It allows expressive musical play through a straightforward link between the selection gestures and excitation or modification of the played sound units, and to be reactive to co-musicians, especially when using live CBCS.

CataRT is a useful and flexible tool that allows composers and musicians to exploit this metaphor as a performance instrument, with a choice of gestural controllers.

For the future, classification of the input gesture could make accessible corresponding classes in the corpus, e.g. by distinguishing attack, sustain, release phases. More advanced gesture analysis and recognition could lead to more expressivity and definition of different playing styles. Finally, machine learning tools for the establishment of adaptive mappings between corpora could increase the usability of audio control.

8. ACKNOWLEDGMENTS

CataRT is essentially based on FTM&Co by Norbert Schnell and collaborators. The work presented here is partially funded by the *Agence Nationale de la Recherche* within the project *Topophonie*, ANR-09-CORD-022, see `http://topophonie.fr`.

9. REFERENCES

[1] A. Bonardi, F. Rousseaux, D. Schwarz, and B. Roadley. La collection numérique comme paradigme de synthèse/composition interactive. *Musimédiane: Revue Audiovisuelle et Multimédia d'Analyse Musicale*, (6), 2011. `http://www.musimediane.com/numero6/COLLECTIONS/`.

[2] C. Cadoz and M. Wanderley. Gesture – Music. In M. Wanderley and M. Battier, editors, *Trends in Gestural Control of Music*. Paris: Ircam, 2000.

[3] V. Johnson and D. Schwarz. Improvising with corpus-based concatenative synthesis. In *(Re)thinking Improvisation: International Sessions on Artistic Research in Music*, Malmö, Sweden, Nov. 2011.

[4] I. Lallemand and D. Schwarz. Interaction-optimized sound database representation. In *DAFx*, Paris, 2011.

[5] D. Schwarz. Concatenative sound synthesis: The early years. *Journal of New Music Research*, 35(1):3–22, Mar. 2006. Special Issue on Audio Mosaicing.

[6] D. Schwarz. Corpus-based concatenative synthesis. *IEEE Sig. Proc. Mag.*, 24(2), Mar. 2007.

[7] D. Schwarz, R. Cahen, and S. Britton. Principles and applications of interactive corpus-based concatenative synthesis. In *Journées d'Informatique Musicale (JIM)*, GMEA, Albi, France, Mar. 2008.

[8] D. Schwarz and N. Schnell. Sound search by content-based navigation in large databases. In *Sound and Music Computing (SMC)*, Porto, July 2009.

[9] P. A. Tremblay and D. Schwarz. Surfing the waves: Live audio mosaicing of an electric bass performance as a corpus browsing interface. In *New Interfaces for Musical Expression*, Sydney, Australia, June 2010.

[10] D. Wessel. Timbre space as a musical control structure. *Computer Music Journal*, 3(2):45–52, 1979.

[11] A. Zils and F. Pachet. Musical Mosaicing. In *Digital Audio Effects (DAFx)*, Limerick, Ireland, Dec. 2001.