Rich Contacts: Corpus-Based Convolution of Audio Contact Gestures for Enhanced Musical Expression

Diemo Schwarz, Pierre Alexandre Tremblay, Alexander Harker

Rich Contacts: Corpus-Based Convolution of Contact Interaction Sound for Enhanced Musical Expression

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

Pierre Alexandre Tremblay
CeReNeM
Huddersfield University, UK
p.a.tremblay@hud.ac.uk

Alexander Harker
CeReNeM
Huddersfield University, UK
ajharker@gmail.com

ABSTRACT

We propose ways of enriching the timbral potential of gestural sonic material captured via piezo or contact microphones, through latency-free convolution of the microphone signal with grains from a sound corpus. This creates a new way to combine the sonic richness of large sound corpora, easily accessible via navigation through a timbral descriptor space, with the intuitive gestural interaction with a surface, captured by any contact microphone. We use convolution to excite the grains from the corpus via the microphone input, capturing the contact interaction sounds, which allows articulation of the corpus by hitting, scratching, or strumming a surface with various parts of the hands or objects. We also show how changes of grains have to be carefully handled, how one can smoothly interpolate between neighbouring grains, and finally evaluate the system against previous attempts.

Keywords

Corpus-Based Concatenative Synthesis, Convolution, Expressivity, Contact Microphones, Audio Gesture

1. INTRODUCTION

Contact gestures are an intuitive way to express musical rhythm and dynamics on any surface or object by hitting it, scratching it, strumming it, etc. It is striking to observe both how easily one can express a dynamic rhythm by table-drumming, one's hands providing a wide range of timbres through the use of fingernails, fingertips, knuckles, and thumb ball, and also how subtle the nuances of timbre and dynamic can be from the onset of the exploration.

The most obvious limiting factor in both sound and variety remains the surface that is hit, and this is where digital musical instruments (DMIs) can contribute to the enrichment of the sonic outcome while keeping the nuances, expressivity, and fully embodied interaction of hand contact gestures, thus turning any surface into an expressive musical instrument.

In this paper, we propose a new way to combine the sonic richness of large corpora of sound (easily accessible via navigation through a space of sound descriptors), with the intuitive gestural interaction on a surface, by capturing the contact interaction sounds with piezo or contact microphones. The captured sound is then convolved with an impulse response (IR) that is, in fact, a sound grain taken from a corpus of sounds such that the grain's timbre and micro-structure is imprinted onto the microphone's signal. Seen the other way around, the grain's timbre is articulated by the spectro-morphology of the contact interaction sound captured by the microphone.

The choice of grain is made by corpus-based concatenative synthesis' content-based approach to navigation of large databases of sound, as implemented, for instance, by the CataRT system[1], where all snippets of sound are laid out in a 2D space according to their sonic characteristics. This 2D space can be navigated with an appropriate 2D controller (mouse, joystick, XY-pad, accelerometer, motion capture) to choose and to smoothly mix grains to be excited by the microphone input.

This paper presents the early conclusions of what is a simple way to join two existing proven technologies in order to multiply respectively their timbral richness and gestural expressivity. It will be evaluated in section 4 as a proof of concept with further ideas for potential development in section 5.

2. RELATED WORK

2.1 Interaction by Contact Gestures

Expressive performance of digital musical instruments (DMIs) via contact microphones on arbitrary surfaces has entered the spotlight through the *Mogees* project[2], where the piezo-source excites physical models of string or bell resonators, allowing to hit, scratch, and strum any object and turn it into a musical instrument. This work was based on research in the real-time music interaction team at Ircam [8, 1], that lead to the *MO* modular musical objects[3], winner of the 2011 Guthman Musical Instrument Competition. The MO software introduces gesture recognition to distinguish different contact gestures (fingertip or -nail scratching, for instance) to then drive different resonators (physical models of strings).

In parallel, Puckette [7] has proposed the use of piezo-captured percussive performance as excitors of nonlinear reverberators, with pre-processing of the piezo signal in order to remove the resonances of the physical system. The paper also makes explicit what is so interesting in keeping the audio signal from the exciter, by opposition to commercial drum triggers in this case, namely:

> *[...] sliding a brush over a drum trigger isn't likely to produce anything useful, whereas doing the same thing on an instrument that operates*

NIME'14, June 30 – July 03, 2014, Goldsmiths, University of London, UK.

[1] http://imtr.ircam.fr/index.php/CataRT
[2] http://www.brunozamborlin.com/mogees
[3] See http://youtu.be/Uhps_U2E9OM?t=1m7s at 1:07.

directly on the audio signal from the contact microphone (as we do here) has the possibility to create a wide range of useful musical sounds. [7]

Independently of the above, the first author has been using piezo pickups on various surfaces in performances since 2009 [10], exploiting the nuances of a corpus according to the sound of impacts which are analysed and mapped to the 2D navigation space of the CataRT software. This last approach uses an attack detector (`bonk~`) that also outputs the spectrum of the attack audio frame. Total energy and centroid of this spectrum is mapped to the x and y target position in the 2D interface to select the grains 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 drawbacks of this method is that the attack detection is not 100% accurate and introduces some latency due to the analysis frame size, but since, in this case, the signal from the piezos was mixed with the audio played by CataRT, the musical interaction still works, as demonstrated in the accompanying video example[4].

2.2 Corpus-Based Concatenative Synthesis

Corpus-based concatenative synthesis (CBCS) systems [9] 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. The selected units are then concatenated (overlapped) and played, after possibly some transformations.

How a musician can interact with and play the corpus as a musical instrument has been the topic of a recent article [10] that shows that the central notion of the interaction is the sound space itself, rather than the method of control. The specific instrument is determined by the controller that steers the navigation, which fall into the groups of positional control, and control by the analysis of audio input.

Two of the authors have also explored ways of using real-time descriptor mapping from a live instrumental source in order to allow expressive multidimensional manipulation of large corpora [12]. That method used multi-dimensional musaicing to transfer performantive nuances of an electric instrument (here a bass guitar) to a corpus of sound grains. While this instrument design was successful with musicians trained on the given instrument and gave very interesting results, it lacks the immediacy and tactility accessible to everyone that a piezo on a table gives. Moreover, the latency and errors in attack detections were quite noticable.

2.3 Convolution

It is important to note that, despite the use of real-time convolution as expressive means having not been explicitely mentioned so far in DMI design literature, convolution as a mode of cross-synthesis has been used by sound designers and computer composers alike for many years, though mostly in deferred time.

In order to allow such real-time exploration, the HISSTools Impulse Response Toolbox has provided the Max community with powerful modular IR manipulation tools [4]. Its modularity has allowed creative research of musical application by practitioners [5]. With a strong musical practice ethic, these tools allow very flexible yet entirely reliable use and abuse. CPU efficient, lightweight, stable and reliable, these pragmatic tools have allowed hands-on methodological research, one of which is showcased here.

[4] https://eprints.hud.ac.uk/20131

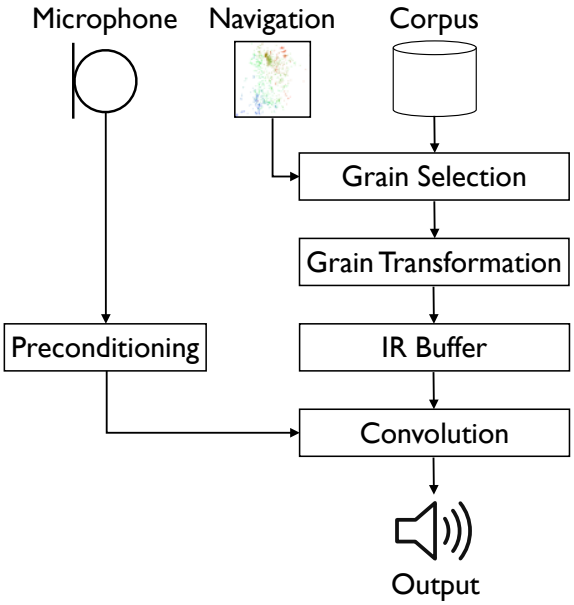

Figure 1: Flowchart of the basic realisation of corpus-based convolution.

3. CORPUS-BASED CONVOLUTION

With the above building blocks available in modular form, the scene was set for integrating the three approaches, (1) intuitive tactile interaction, (2) descriptor-based corpus navigation, (3) real-time convolution as a means of cross-synthesis, with the hope of avoiding the limitations of any one of these alone. In other words, there was a desire for timbral richeness, reuse of virtuosity, and expressive gestural control, that was made possible by a slight abuse of convolution, hijacking it from its usual filtering and room acoustics contexts.

This section will explain the principle and implementation of corpus-based convolution (section 3.1) and then tackle two problems resulting from its usage in an interactive DMI: firstly, how to avoid abrupt changes and clicks when grains enter and leave the convolver (section 3.2), and secondly how to make the scattered bits of sound in the corpus into a smooth map through which the player can navigate (section 3.3). Preprocessing of the piezo input signal to enhance the sonic outcome is then tackled (section 3.4).

3.1 Basic Realisation

The basic realisation of corpus-based convolution is achieved by making a bridge from a selection within the granular corpus-based concatenative synthesis navigator to a convolution engine via an audio buffer, as illustrated in figure 1. Every time the player navigates in the descriptor space (for instance in a 2D representation as in figure 2) such that a new grain becomes closest, that grain is taken as an impulse response to be convolved with the microphone input signal. In its simplest form, the navigation is done by moving a 2D target via a positional controller such as a mouse, graphics tablet, or touch screen (cf. [10], positional control). Note that the grain which is used as IR can be processed like any

grain by the usual time-domain processes, such as transposition, gain, length and envelope change, reversal, all with possible random variations. This further enhances the sonic richness and expressive capabilities of the system.

Technically, in our prototyping system based on CataRT, FTM, and its Gabor library, the `catart.synthesis` module outputs the grain as a 1-column matrix of floating-point numbers (`fmat`). The matrix is copied into a `buffer~` object via the `ftm.buffer` object. That `buffer~` is referenced by a `multiconvolve~` object from the HISSTools which is notified of the update via a message.

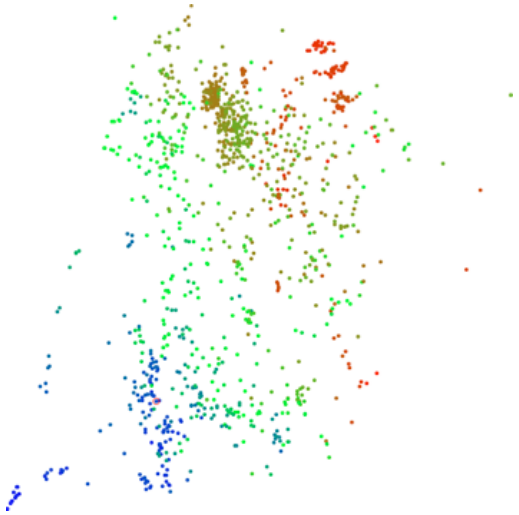

Figure 2: Example of a corpus projected to 2D for efficient navigation, plotted by Spectral Centroid (x), Periodicity (y), NoteNumber (colour).

The basic realisation above captures the essence of the interaction that allows to choose and then articulate one grain by contact gestures, however it is not yet sufficient for a varied musical performance. The next three sections will extend the basic realisation to enrich further the sonic outcome.

3.2 Handling Grain Changes

When a new grain is selected by navigation in the descriptor space, it is set as a new impulse response for the `multiconvolve~` convolver. This means that from that moment on, the output signal ceases to be the convolution with the old grain, but the one with the new grain, which leads to an abrupt timbral change or even a click.

The aim is to let the old grain "finish", while the new grain is already being convolved, in effect reproducing an overlap-add mechanism. This can be achieved by multiplying the convolution modules by means of Max/MSP's `poly~` multi-voice object. The old grain's input signal is faded out during a release time, here 200 ms, which allows the old grain to die out. Note that the convolution output can still continue for the duration of the grain, so only after $t_r = releasetime + grainduration$ can the voice be freed. Simultaneously, a new convolution voice is allocated and receives the new grain and in faded-in by a short attack ramp of 10 ms.

We need a polyphony higher than two, since during the fade out time of one grain, again a new grain can be selected, when the navigation speed in the corpus is fast, or the grains are very dense such that a trajectory crosses several grains within t_r. In order not to tax too much the CPU power of the DMI, we have found out through empirical testing that a polyphony of 8 streams was a sufficient compromise between fluid crossfading and CPU usage.

3.3 Smooth Interpolation Between Grains

In the basic realisation so far, the navigation in the corpus will switch between impulse response grains as soon as the target position moves closest to a new grain. This means the 2D navigation interface is split into discrete Voronoi-polygons.

While for some musical aim this might be appropriate, for instance the creation of rapidly switching timbre transitions, the continuous nature of the navigation space suggests a smoother way to transition between timbres given by the grains in the corpus.

One approach to achieve smooth timbre transitions is to interpolate between the 3 surrounding points in the navigation space. The relative distances to these points can then be used to derive a weight or amplitude factor [3] to mix the convolutions with the 3 grains. We use this simplified formula to calculate gain factors g_i in dB for $i = 1..3$ convolutions, based on the distances d_i to the 3 closest points:

$$g_i = \frac{g_{min}\, d_i}{d_{min} \sum_{i=1}^{3} d_i} \tag{1}$$

where $g_{min} = -64$ dB is the minimum gain for the maximum distance $d_{min} = \frac{2}{3}$. The gain for a grain becomes maximum (0 dB) when its position coincides with the target position. In the middle of a triangle, all surrounding grains have equal gain.

The implementation needs 3 channels of the 8-voice polyphonic convolvers above whose output signals are attenuated by g_i and mixed.

3.4 Pre-Emphasis

As the direct signal from the piezo is used to convolve the grains, preprocessing of its sonic characteristics is required: indeed the microphone will exhibit some grain of the surface on which it is applied, most of the time resulting in a dull top-end, with overdynamic spikes when getting too close to the microphone. Moreover, the piezo's high impedance and low gain usually requires high pre-amplifier gains, which is prone to generate background noise.

We have found that a three band filtering (low high-pass, a band-cut on the main resonance and a boosting high-shelf), followed by a subtle low-threshold expander and a high-threshold limiter, is a suitable pre-conditioning for the signal and yield results that are much closer to the timbral characteristics of the IRs selected. By opposition to the work by Puckette [7], we did not want to lose the filtering quality of the performing surface captured in the piezo signal as we considered that this filter is part of the instrument's sound. Nevertheless, the modularity of the current implementation would easily allow the pre-emphasis filtering to be replaced by a FIR-filter extrapolated from inverted surface resonance, as proposed by Puckette [7].

4. EARLY EVALUATION

As this paper is a proof of concept, the authors are eager to share the results in a usable manner, allowing the DMI designers and users to experience the results for themselves. The code is therefore available with all its dependencies[4]. What follows is a short yet promising early evaluation.

When the new system is played in comparison with the previous `bonk~`-based instance, the first striking improvement is the immediacy of the instrument: the excitation of the grains with the piezo sound's complex gestures is direct, rich, palpable, without any noticable latency. A complex dynamics between both hands, namely the percusive

one and the corpus-navigator one, quickly establishes itself and allows many games: with more percussive grains, the piezo becomes a filter, adding texture; with more percussive piezo gestures, the grains pop to life in a very rich way, while retaining their temporal micro-structure.

In order to illustrate the difference, the reader is invited to download the instrument. For the reader who would not want to run the code, a short comparison film has been made available[4], as well as further sound examples of this new instrument. In the first video, we can hear the same gesture performed by itself, then on the old `bonk~`-based DMI, and finally on the new DMI, first without and then with pre-emphasis. As the examples show, the rich palette of nuances of the piezo input are cast upon the grains, giving it more depth and spectral contour. Moreover, the absence of latency makes the new DMI more immediate than the previous one.

Compared to the bass guitar musaicing [12] described in section 2.2, what stands out again is the new method's absence of latency, as well as its ability to keep a wide range of nuances, as produced by the complex hybridisation of the signals. Moreover, it confirms and solves the two issues of the musaicing instrument, namely the granularity of the sound and the attack problems: in the new method, it seems that the complexity of the signal combination mitigates the granular effect, giving it a much smoother articulation.

5. CONCLUSIONS AND FUTURE WORK

This proof-of-concept implementation of corpus-based convolution efficiently combines the intuitive and expressive control of dynamics and temporal shape provided by hand-percussion playing on a surface, with the rich and varied timbres chosen from large corpora of sound by content-based selection.

One significant part of the digital musical interface has not been treated specifically here, namely how to control the navigation within the corpus of potential IRs. Currently, it is kept in its already existing form of 2D control navigation as described in section 1. This means however that one hand will be used for the position input and only one hand is left for contact interaction.

With the early results presented here being so promising, we look forward to future developments of the navigation control by other means in order to have both hands free to play the contact microphones, to push further this approach:

Hand Position Tracking either by 2D or 3D camera, or even audio triangulation, would perfectly integrate selection of grains by the position of the hands and their expressive play on a surface. Here, the interaction space should be optimised by evenly distributing grains over the whole playing surface [6].

Audio Mapping similar to that used in the previous trigger-based system using `bonk~` would extrapolate the dynamics and timbre of the contact interaction sound to not only articulate a grain but also change the selection.

Voice Control would use the performer's voice timbre and morphology to select grains by similarity matching combined with an adaptive projection that maps the space of possible voice timbres to a given corpus' descriptor space [11, 2].

Pedals could also be used, or a controller mat, in the spirit of guitar players extending their expressive power by using their free limbs.

Other Sources than a piezo could be used. A hybrid project between electric instrument control and certain descriptor mapping is also full of potential to be explored to solve the problems of latency and granularity mentioned in [12].

As this exploration is in its early days, there are also many potential optimisations to be implemented: First, we could unify the gain changes due to distance-based mixing and due to the handling of the release by incorporating both into the amplification of the input signal before the convolution. Second, the partitioned convolution algorithm used in HISSTools multiplies the FFT spectra of the input signal and the impulse response to carry out the convolution more efficiently. If we now constitute the impulse response by interpolating between the 3 closest grains in the spectral domain, we only need one convolution instead of 3 separate ones. Spectral domain representations of the grains could even be precomputed for the whole corpus.

6. REFERENCES

[1] F. Bevilacqua, N. Schnell, N. Rasamimanana, B. Zamborlin, and F. Guédy. Online gesture analysis and control of audio processing. In Solis and Ng, editors, *Musical Robots and Interactive Multimodal Systems: Springer Tracts in Advanced Robotics*, volume 74, pages 127–142. Springer Verlag, 2011.

[2] S. Fasciani and L. Wyse. One at a Time by Voice: Performing with the Voice-Controlled Interface for Digital Musical Instruments. In *SI 13: NTU/ADM Symposium on Sound and Interactivity*, Singapore, 2013.

[3] A. Freed, J. MacCallum, A. Schmeder, and D. Wessel. Visualizations and Interaction Strategies for Hybridization Interfaces. In *Proc. NIME*, 2010.

[4] A. Harker and P. A. Tremblay. The HISSTools impulse response toolbox: Convolution for the masses. In *Proc. ICMC*, pages 148–155. 2012.

[5] A. Harker and P. A. Tremblay. Rethinking the box: Approaches to the reality of electronic music performance. In *IRCAM Forum*, 2013[5].

[6] I. Lallemand and D. Schwarz. Interaction-optimized sound database representation. In *Digital Audio Effects (DAFx)*, Paris, France, 2011.

[7] M. Puckette. Infuriating nonlinear reverberator. In *Proc. ICMC*, Huddersfield, UK, 2011.

[8] N. Rasamimanana, F. Bevilacqua, N. Schnell, F. Guedy, E. F. Come Maestracci, B. Zamborlin, J.-L. Frechin, and U. Petrevsky. Modular musical objects towards embodied control of digital music. In *Tangible Embedded and Embodied Interaction*, 2011.

[9] D. Schwarz. Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, 24(2), 2007.

[10] D. Schwarz. The Sound Space as Musical Instrument : Playing Corpus-Based Concatenative Synthesis. In *Proc. NIME*, Ann Arbor, MI, USA, 2012.

[11] D. Stowell and M. Plumbley. Timbre remapping through a regression-tree technique. In *Sound and Music Computing (SMC)*, 2010.

[12] P. A. Tremblay and D. Schwarz. Surfing the waves: Live audio mosaicing of an electric bass performance as a corpus browsing interface. In *Proc. NIME*, pages 447–450, Sydney, Australia, 2010.

[5] See also the last presentation of Nov. 22nd at http://forumnet.ircam.fr/forum-workshops-2013-videos/