Real-Time Corpus-Based Concatenative Synthesis with CataRT

Diemo Schwarz, Grégory Beller, Bruno Verbrugghe, Sam Britton

Ircam – Centre Pompidou, Paris, France

(schwarz|beller|verbrugg|britton)@ircam.fr

ABSTRACT

The concatenative real-time sound synthesis system *CataRT* plays grains from a large corpus of segmented and descriptor-analysed sounds according to proximity to a target position in the descriptor space. This can be seen as a content-based extension to granular synthesis providing direct access to specific sound characteristics. *CataRT* is implemented in *Max/MSP* using the FTM library and an SQL database. Segmentation and MPEG-7 descriptors are loaded from SDIF files or generated on-the-fly. *CataRT* allows to explore the corpus interactively or via a target sequencer, to resynthesise an audio file or live input with the source sounds, or to experiment with expressive speech synthesis and gestural control.

1. INTRODUCTION

Corpus-based concatenative synthesis methods are attracting more and more interest in the musical sound synthesis and content-based processing communities. They use a large database of source sounds, segmented into *units*, and a *unit selection* algorithm that finds the sequence of units that match best the sound or phrase to be synthesised, called the *target*.

The selection is performed according to the *descriptors* of the units, which are characteristics extracted from the source sounds, or higher level descriptors attributed to them. The selected units can then be transformed to fully match the target specification, and are concatenated.

These methods allow various applications, such as high level instrument synthesis, resynthesis of audio, also called *mosaicing*, texture and ambience synthesis, artistic speech synthesis, and interactive explorative synthesis in different variants, which is the main application of the *CataRT* synthesis system.

Explorative real-time synthesis from heterogeneous sound databases allows to exploit the richness of detail of recorded sound while retaining efficient control of the acoustic result by using perceptually meaningful descriptors to specify a target in the multi-dimensional descriptor space. If the selection happens in real-time, this allows to browse and explore a corpus of sounds interactively.

CataRT[1] is a collection of patches for *Max/MSP* using the FTM, *Gabor*, and MnM extensions[2]. It is released as free open source software under the GNU general public license (GPL)

2. RELATED WORK

CataRT is based on the previous developments *Caterpillar* for non real-time data-driven concatenative musical sound synthesis [1, 2], and *Talkapillar* for artistic text-to-speech synthesis, hybrid concatenation between music and speech, and descriptor-driven or context-sensitive voice effects [3]. In these systems, the unit selection algorithm is by Viterbi path-search, which finds the globally optimal sequence of database units that best match the given target units using two cost functions: The *target cost* expresses the similarity of a target unit to the database units by weighted Euclidean distance, including a context around the target. The *concatenation cost* predicts the quality of the join of two database units by join-point continuity of selected descriptors.

Many other approaches to concatenative corpus-based synthesis appeared over the last few years. They are compared and classified in [4].

2.1. Real-Time Concatenative Synthesis

Contrary to the systems above, real-time concatenative synthesis systems can not provide a globally optimal selection, because the target is not known entirely in advance. Also, concatenation quality is rarely included in the selection algorithm. They fall into two groups: The first group matches low- or high-level descriptors of the corpus units to target descriptors (*MoSievius* [5], *Synful* [6], *Ringomatic* [7]). The second group performs a spectral match based on FFT-frames where the target is given by live audio input. (*Input Driven Resynthesis* [8], *Scrambled Hackz*[3]), or on spectro-temporal "sound lexemes" (*SoundSpotter* [9]).

2.2. Granular Synthesis

One source of inspiration of the present work is granular synthesis, which takes short snippets (*grains*) out of a sound file, at an arbitrary rate. These grains are played back with a possibly changed pitch, envelope, and volume. The position and length of the snippets are controlled interactively, allowing to scan through the soundfile, in any speed. Granular synthesis is rudimentarily corpus-based, considering that there is no analysis, the unit size is determined arbitrarily, and the selection is limited to choosing the position in one single sound file. However, its concept of exploring a sound interactively, when combined with a pre-analysis of the data and thus enriched by a targeted selection, results in a precise control over the output sound characteristics, as realised in *CataRT*.

3. MODEL

CataRT's model is a multidimensional space of descriptors, populated by the sound units. The user controls a target point in a lower-dimensional projection of that space with a selection radius around it, and the selection algorithm selects the units closest to the target or within the radius. The actual triggering of the unit is independent of the selection and can happen at any rate.

1. http://concatenative.net
2. http://www.ircam.fr/ftm
3. http://www.popmodernism.org/scrambledhackz/

3.1. Analysis

To get data into the corpus, either all data (audio, segment markers, raw descriptors) are loaded from preanalysed files, or the descriptors are analysed in *CataRT* and an additional onset detection stage segments the sound files. At this point, all analysis takes place inside *CataRT*, which means that we could just as well use real-time audio input that is segmented and analysed on the fly, to feed the corpus. The audio could come, for example, from a musician on stage, the last several minutes of whose playing constitutes the corpus from which a laptop improviser selects units.

Markers generated externally can be loaded from SDIF or ASCII files; internal segmentation calculation is either by arbitrary grain segmentation, by split according to silence (given a threshold), by high-frequency content, or by transient analysis. There is also a mode where imported sound files are taken as a whole, which is appropriate for sets of drum and percussion sounds.

Descriptors are either imported from precalculated MPEG-7 low-level descriptor files, or calculated in the patch. Details for the 230 imported MPEG-7 signal, perceptual, spectral, and harmonic descriptors can be found in [2], following the definitions from [10].

The descriptors calculated in the patch in batch mode, i.e. faster than real-time, thanks to *Gabor*'s event-based signal frame processing, are the fundamental frequency, aperiodicity, and loudness found by the *yin* algorithm [11], and a number of spectral descriptors from [12]: spectral centroid, sharpness, spectral flatness, high and mid frequency energy, high frequency content, first order autocorrelation coefficient (that expresses spectral tilt), and energy.

Note that also descriptors describing the units' segments themselves, such as the unit's unique id, the start and end time, its duration, and the soundfile it came from, are stored. It is convenient to duplicate this information as descriptors to make it available for selection.

The time-varying raw descriptors at FFT-frame rate have to be condensed to a fixed number of scalar values to characterise a unit. These *characteristic values* [2] express the general evolution over time of a descriptor with its mean value, variance, slope, curve, min, max, and range.

3.2. Data

Data is kept in FTM data structures such as tables, dictionaries, matrices. The unit descriptor data is kept in one big matrix with one column per descriptor and one unit per row. For persistent storage of corpora, a layer around the relational database management system *SQLite* keeps track of soundfiles, segments, and unit descriptor data. The Sound Description Interchange Format (SDIF)[4] is used for well-defined interchange of data with external descriptor analysis and segmentation programs.

3.3. Selection

Because of the real-time orientation of *CataRT*, we cannot use the globally optimal path-search style unit selection based on a Viterbi algorithm as in *Caterpillar*, neither do we consider concatenation quality, for the moment. Instead, the selection is based on finding the units closest to the current position x in the descriptor space, in a geometric sense, i.e. on appropriately scaled dimensions: A straightforward way of achieving this is to calculate the Mahalanobis distance d between x and all units, i.e. the distance normalised by the standard deviation of each chosen descriptor over the corpus. Either the unit with minimal d is selected, or one randomly chosen from the set of units with $d < r^2$, when a selection radius r is specified.

To improve the efficiency of selection, the units in the descriptor space are indexed by an optimised multi-dimensional k-nearest neighbour index. The algorithm described in [13] constructs a search tree by splitting up the descriptor space along the hyperplane perpendicular to the principal component vector, and thus achieving a maximal separation of units. This is then repeated for each sub-space until only a few units are left in each leaf node of the resulting tree. The k-nearest neighbour search can then, at each step down the tree, eliminate approximately half of the units, by just one distance calculation with the subspace boundary.

3.4. Synthesis

CataRT's synthesis applies a choosable short fade-in and fade-out to the sound data of a selected unit, which is then pasted into the output delay-line buffer, possibly with a random delay. Other manipulations similar to a granular synthesis engine can be applied: the copied length of the sound data can be arbitrarily changed (de facto falsifying the selection criteria) to achieve granular-style effects or clouds of overlapping grains. Also, changes in pitch by resampling and loudness changes are possible. Note that, because the actual pitch and loudness values of a unit are known in its descriptors, it is possible to specify precise pitch and loudness values that are to be met by the transformation.

However, granular synthesis is only one possible realisation of synthesis. Other components might store the sound data in spectral or additive sinusoidal representations for easier transformation and concatenation.

3.5. Interface

Because displaying and navigating in a high-dimensional space is not practical, the descriptor space is reduced to a 2-dimensional projection according to two selectable descriptors.

CataRT's user interface follows the model–view–controller (MVC) design principle. The view (figure 1) plots a 2-dimensional projection of the units in the descriptor space plus a 3rd descriptor being expressed on a colour scale. Note that that the display is dynamic, i.e. multiple views can be instantiated that can connect to the same data component, or one view can be switched between several data instances, i.e. corpora.

In these displays, the mouse serves to move the target point in the descriptor space. Additional control possibilities are MIDI input from fader boxes to set more than two descriptor target values and limit a selectable descriptor range, and advanced input devices for gestural control (see section 4.2).

Independent of the current position, several sources for triggering playing of the currently closest unit exist: an obvious but quite interesting mode plays a unit whenever the mouse moves. De-facto, the friction of the mouse provides an appropriate force-feedback, so that this mode is called *bow*. To avoid the strident repetitions of units, the mode *fence* plays a unit whenever a different unit becomes the closest one (named in homage to clattering a stick along a garden fence). The *beat* mode triggers units regularly via a metronome, and the *chain* mode triggers a new unit whenever the previous unit has finished playing.

4. http://www.ircam.fr/sdif

CataRT incorporates a basic loop-sequencer that allows to automate the target descriptor control (see figure 2). Also the evolution of the weight for a descriptor can be sequenced (second curve in figure 2), such that at the desired times, the target descriptor value is enforced, while at others the selection is less dependent on this descriptor.

4. APPLICATIONS

4.1. Explorative Granular Synthesis

The principal application of *CataRT* is the interactive explorative synthesis from a sound corpus, based on musically meaningful descriptors. Here, granular synthesis is extended by a targeted selection according to the content of the sound base. One could see this as abolishing the temporal dimension of a sound file, and navigating through it based on content alone.

Usually, the units group around several clusters. With corpora with mixed sources, such as environmental noises, voice, and synthetic sounds, interesting overlaps in the descriptor space occur and can be exploited. Figure 1 shows the *CataRT* main patch with an example of a corpus to explore.

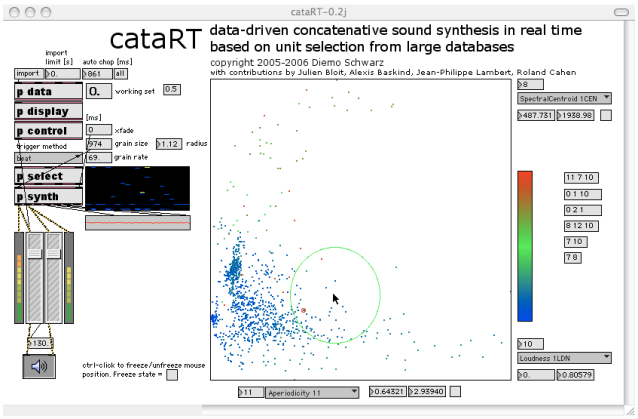

Figure 1: Screen shot of the explorative synthesis interface

4.2. Gesture-Controlled Synthesis

The present model of controlling the exploration of the sound space by mouse position is of course very basic and limited. That's why we're working on the integration of a general mapping layer for gestural control into *CataRT*. This allows more expressive musical interaction and tapping into more than just two dimensions by analysis of higher level gestural parameters from more advanced input devices such as graphics tablets. Clustering, rotation and warping of the descriptor space (by multi-dimensional scaling or magnifying-glass type transformations) maximises the efficiency of the interaction, leading to greater expressivity. Sub-spaces of units can be selected by navigation or according to the temporal evolution of the control gesture (attack, sustain, release phases can have their own sub-space to select units from). Note that each sub-space can have its own projection to the most pertinent descriptors therein.

4.3. Audio-Controlled Synthesis

As all the analysis can take place in *CataRT* itself, it is possible to analyse an incoming audio signal for descriptors and use these to control the synthesis, effectively resynthesising a signal with sounds from the database. The target descriptor evolution can also be derived from a sound file, by analysing and segmenting it and playing its controlling descriptors from the sequencer.

CataRT is used as a compositional and orchestration tool in the context of a piece for banjo and electronics by Sam Britton. The work takes large databases of recorded instrumental improvisations and uses concatenative synthesis to re-sequence and orchestrate these sequences. In this context, the concatenation process acts as a kind of oral catalyst, experimentally re-combining events into harmonic, melodic and timbral structures, simultaneously proposing novel combinations and evolutions of the source material.

4.4. Data-Driven Drumbox

A slighly more rigid variation of the *CataRT* sequencer splits the target descriptor sequence into a fixed number of beats, where, for each beat, one sound class can be chosen. The selection within each soundclass, however, is gouverned by a freely editable descriptor curve, or real-time control.

4.5. Expressive Speech Synthesis

Research in expressive speech synthesis [14] takes advantage of real-time content-based manipulation and synthesis. Special descriptors are used for the speech units, e.g. acoustic descriptors like formant frequencies and pitch and symbolic descriptors such as the phonemes and the grammar. Here, a sentence is set up as a target sequence of successive segments of speech sounds and their descriptors. We can then change descriptor values of the target sequence to catch other sound units (speech or other) from the database. Figure 2 shows the pitch and the phonemic transcription of a sequence of units corresponding to the utterance "au revoir". It can be seen that we have graphically modified pitch around the temporal index shown by the cursor. All descriptors can be modified in such an easy way to change locally the selection of a unit within a sequence. Text can also be rewritten using the keyboard as in the example where we have delayed the last "R".

There are two different ways to define temporal boundaries of the speech segments: The first is by segmental units like diphones, phones or semiphones as presented in figure 2, and leads to a small number of units per sentence which allows to use a many files to provide classical TTS with the capability to draw the prosody. The second is by pitch synchronous segmentation in periods of the speech signal. A large number of units per sentence are created which permits a fine grained selection. In this case, modifications of the descriptor sequence lead to a progressive morphing between spoken and singing synthesis, for instance. It is also possible to add jitter (perturbation of the pitch) to provide expressive speech synthesis.

5. FUTURE WORK

Except the obvious work on adding more descriptors and improving the visualisation, interesting questions of control and interaction are raised by the present work.

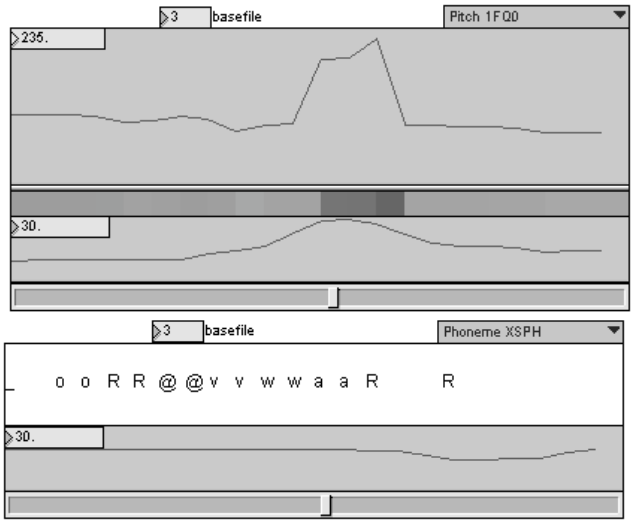

Figure 2: The *CataRT* descriptor sequencer

The used corpora are in general unevenly distributed over the descriptor space. Many units are concentrated in clusters, whereas large parts of the space are relatively empty. This is first a problem of interaction and visualisation, which should allow zooming into a cluster to navigate through the fine differences within. However, the model of navigating through the descriptor space could be refined by a notion of subspaces with links to other subspaces. Note that, within different clusters, possibly different descriptors express best the intra-cluster variation.

CataRT should take care of concatenation, at least in a limited way, by considering the transition from the previously selected unit to the next one. The concatenation cost could be given by descriptor continuity constraints, spectral distance measures, or by a precalculated distance matrix, which would also allow distances to be applied to symbolic descriptors such as phoneme class. The concatenation distance could be derived from an analysis of the corpus:

It should be possible to exploit the data in the corpus to analyse the natural behaviour of an underlying instrument or sound generation process. By modeling the probabilities to go from one cluster of units to the next, we would favour the typical articulations of the corpus, or, the synthesis left running freely would generate a sequence of units that recreates the texture of the source sounds.

To make more existing sound collections available to *CataRT*, an interface to the *Caterpillar* database, to the *freesound* repository, and other sound databases is planned. The *freesound* project,[5] is a collaboratively built up online database of samples under licensing terms less restrictive than the standard copyright, as provided by the *Creative Commons* family of licenses. A transparent net access from *CataRT* to this sound database, with its 170 unit descriptors already calculated, would give us an endless supply of fresh sound material.

6. CONCLUSION

We presented the *CataRT* system that implements a new model of interactive exploration of, and navigation through a sound corpus. The concatenative synthesis approach is a natural extension of granular synthesis, augmented by content-based selection and control, but keeping the richness of the source sounds. By its real-time analysis capabilities, the system can also perform context-based effects. The applications are only beginning to fathom all the possibilities this model allows for interaction modes and visualisation. Because of their object-oriented software architecture, the *CataRT* components can serve as a toolbox to build individual applications.

7. REFERENCES

[1] Diemo Schwarz, "A system for data-driven concatenative sound synthesis," in *DAFx*, Verona, 2000.

[2] Diemo Schwarz, *Data-Driven Concatenative Sound Synthesis*, Thèse de doctorat, Université Paris 6 – Pierre et Marie Curie, Paris, 2004.

[3] Grégory Beller, Diemo Schwarz, Thomas Hueber, and Xavier Rodet, "A hybrid concatenative synthesis system on the intersection of music and speech," in *Journées d'Informatique Musicale (JIM)*, MSH Paris Nord, St. Denis, France, June 2005.

[4] Diemo Schwarz, "Concatenative sound synthesis: The early years," *Journal of New Music Research*, 2006, Special Issue on Audio Mosaicing: Feature-Driven Audio Editing/Synthesis.

[5] Ari Lazier and Perry Cook, "MOSIEVIUS: Feature driven interactive audio mosaicing," in *DAFx*, London, 2003.

[6] Eric Lindemann, "Musical synthesizer capable of expressive phrasing," US Patent 6,316,710, Nov. 2001.

[7] Jean-Julien Aucouturier and Franois Pachet, "Ringomatic: A Real-Time Interactive Drummer Using Constraint-Satisfaction and Drum Sound Descriptors," in *International Symposium on Music Information Retrieval (ISMIR)*, London, UK, 2005.

[8] M. Puckette, "Low-dimensional parameter mapping using spectral envelopes," in *Proc. ICMC*, Miami, 2004.

[9] Michael Casey, "Acoustic lexemes for organizing internet audio," *Contemporary Music Review*, vol. 24, no. 6, 2005.

[10] Geoffroy Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Tech. Rep., Ircam – Centre Pompidou, 2004.

[11] Alain de Cheveigné and Nathalie Henrich, "Fundamental Frequency Estimation of Musical Sounds," *Journal of the Acoustical Society of America (JASA)*, 2002.

[12] Julien Bloit, "Analyse temps réel de la voix pour le contrôle de synthèse audio," Master-2/SAR ATIAM, UPMC (Paris 6), Paris, 2005.

[13] Wim D'haes, Dirk van Dyck, and Xavier Rodet, "PCA-based branch and bound search algorithms for computing K nearest neighbors," *Pattern Recognition Letters*, vol. 24, no. 9–10, pp. 1437–1451, 2003.

[14] Grégory Beller, Thomas Hueber, Diemo Schwarz, and Xavier Rodet, "Speech rates in french expressive speech," in *Speech Prosody*, Dresden, Germany, 2006.

5. http://iua-freesound.upf.es