CURRENT RESEARCH IN CONCATENATIVE SOUND SYNTHESIS

Diemo Schwarz
Ircam – Centre Pompidou
1, place Igor-Stravinsky, 75004 Paris, France

ABSTRACT

Concatenative synthesis is a promising method of musical sound synthesis with a steady stream of work and publications in recent years. It uses a large database of sound snippets to assemble a given target phrase. We explain its principle and components and its main applications, and compare existing concatenative synthesis approaches. We then list the most urgent problems for further work on concatenative synthesis.

1. INTRODUCTION

Concatenative synthesis methods 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. However, if the database is sufficiently large, the probability is high that a matching unit will be found, so the need to apply transformations is reduced. The units can be *non-uniform*, i.e. they can comprise a sound snippet, an instrument note, up to a whole phrase.

Concatenative synthesis can be more or less *data-driven*, where, instead of supplying rules constructed by careful thinking as in a *rule based approach*, the rules are induced from the data itself. The advantage of this approach is that the information contained in the many sound examples in the database can be exploited.

The current work on concatenative sound synthesis (CSS) focuses on three main applications:

High Level Instrument Synthesis Because concatenative synthesis is aware of the context of the database as well as the target units, it can synthesise natural sounding transitions by selecting units from matching contexts. Information attributed to the source sounds can be exploited for unit selection, which allows high-level control of synthesis, where the fine details lacking in the target specification are filled in by the units in the database.

Resynthesis of audio A sound or phrase is taken as the audio score, which is resynthesized with the same pitch, amplitude, and timbre characteristics using units from the database.

Free synthesis from heterogeneous sound databases offers a sound composer efficient control of the result by using perceptually meaningful descriptors, and allows to browse and explore a corpus of sounds.

Concatenative synthesis sprung up independently in multiple places and is sometimes referred to as *mosaicing*. It is a complex method that needs many different concepts working together, thus much work on only one single aspect fails to relate to the whole. It has seen accelerating development over the past few years as can be seen in the chronology in section 3. There is now the first commercial product available (3.13), and, last but not least, ICMC 2004 saw the first musical pieces using concatenative synthesis (3.12). We try in this article to acknowledge this young field of musical sound synthesis that has been identified as such only five years ago.

Any CSS system must perform the following tasks, sometimes implicitly. This list of tasks will serve later for a taxonomy of existing systems.

Analysis The source sound files are segmented into units and analysed to express their characteristics with sound descriptors. Segmentation can be by automatic alignment of music with its score for instrument corpora, by blind or arbitrary grain segmentation for free and re-synthesis, or can happen on-the-fly. The descriptors can be categorical (class membership), static (constant over a unit), or dynamic (varying over the duration of a unit).

Database Source file references, units and unit descriptors are stored in a database. The subset of the database that is preselected for one particular synthesis is called the *corpus*.

Target The target specification is generated from a symbolic score (expressed in notes or descriptors), or analysed from an audio score (using the same segmentation and analysis methods as for the source sounds).

Selection Units are selected from the database that match best the given target descriptors according to a distance function and a concatenation quality function. The selection can be local (the best match for each target unit is found individually), or global (the sequence with the least total distance if found).

Synthesis is done by concatenation of selected units, possibly applying transformations. Depending on the application, the selected units are placed at the times given by the target (musical or rhythmic synthesis), or are concatenated with their natural duration (free synthesis or speech synthesis).

2. RELATED WORK

Concatenative synthesis is at the intersection of many fields of research, such as music information retrieval, database technology, real-time and interactive methods, sound synthesis models, musical modeling, classification, perception. Concatenative *text-to-speech* synthesis shares many concepts and methods with concatenative sound synthesis, but has different goals. *Singing voice synthesis* occupies an intermediate position between speech and sound synthesis and often uses concatenative methods. For instance, Meron [13] uses an automatically constituted large unit database of one singer.

Content based processing is a new paradigm in digital audio processing that shares the analysis with CSS. It performs symbolic manipulations of elements of a sound, rather than using signal processing alone. Lindsay [12] proposes context-sensitive effects by utilising MPEG-7 descriptors. Jehan [7] works on the objet segmentation and perception-based description of audio material and then manipulates the audio in terms of its musical structure. The Song Sampler [1] is a system which automatically samples meaningful units of a song, assigns them to the keys of a Midi-keyboard to be played with by a user.

Related to selection based sound synthesis is *music selection* where a sequence of songs is generated according to their characteristics and a desired evolution over the playlist. An innovative solution based on constraint satisfaction is proposed in [16], which ultimately inspired the use of constraints for CSS in [27] (section 3.3).

The *Musescape* music browser [25] works by specifying high-level musical features (tempo, genre, year) on sliders. The system then selects in real time musical excerpts that match the desired features.

3. CHRONOLOGY

Approaches to musical sound synthesis that are somehow data-driven and concatenative can be found throughout history. They are usually not identified as such but can be arguably seen as instances of fixed inventory or manual concatenative synthesis.

The groundwork for concatenative synthesis was laid in 1948 by the *Groupe de Recherche Musicale* (GRM) of Pierre Schaeffer, using for the first time recorded segments of sound to create their pieces of *Musique Concrète*. Schaeffer defines the notion of *sound object*, which is a clearly delimited segment in a source recording. This is not so far from what is here called *unit*.

Concatenative aspects can also be found in *sampling*. The sound database consists of a fixed unit inventory analysed by instrument, playing style, pitch, and dynamics, and the selection is reduced to a fixed mapping of Midi-note and velocity to a sample. Similarly, when choosing *drum loops* from sampling CDs, a dance music composer is implicitly performing a selection from a large amount of data, guided by their characteristics.

Granular synthesis can be seen as rudimentarily data-driven, but there is no analysis, the units size is determined arbitrarily, and the selection is limited to choosing the position in one sound file. However, its concept of exploring a sound interactively could be combined with a pre-analysis of the data and thus enriched by a targeted selection and the resulting control over the output sound characteristics, i.e. where to pick the grains that satisfy the wanted sound characteristics.

3.1. Plunderphonics (1993)

Plunderphonics [15] is John Oswald's artistic project consisting of songs made up from tens of thousands of snippets from a decade of pop songs, selected and assembled by hand. The sound base was manually labeled with musical genre and tempo as descriptors.

3.2. Caterpillar (2000)

Caterpillar [19, 20, 21, 22], performs data-driven concatenative musical sound synthesis from large heterogeneous sound databases. Units are segmented by automatic alignment of music with its score [14], or by blind segmentation. The descriptors are based on the MPEG-7 low-level descriptor set [17], plus descriptors derived from the score and the sound class. The low-level descriptors are condensed to unit descriptors by modeling of their temporal evolution over the unit (mean value, slope, range, etc.) The database is implemented using a relational SQL database management system for reliability and flexibility.

The unit selection algorithm inspired from speech synthesis finds the sequence of database units that best match the given synthesis target units using two cost functions: The *target cost* expresses the similarity of a target unit to the database units, including a context around the target, and the *concatenation cost* predicts the quality of the join of two database units. The optimal sequence of units is found by a Viterbi algorithm as the best path through the network of database units.

The *Caterpillar* framework is also used for expressive speech synthesis [2], and first attempts for hybrid synthesis combining music and speech are described in [3].

3.3. Musical Mosaicing (2001)

Musical Mosaicing, or *Musaicing* [27], performs a kind of automated remix of songs. It is aimed at a sound database of pop music, selecting pre-analysed homogeneous snippets of songs and reassembling them. Its great innovation was to formulate the unit selection as a constraint solving problem. The set of descriptors used for the selection is: mean pitch, loudness, percussivity, timbre. Work on adding more descriptors has picked up again with [28].

3.4. Soundmosaic (2001)

Soundmosaic [5] constructs an approximation of one sound out of small units of varying size from other sounds. For version 1.0 of *Soundmosaic*, the selection of the best source unit uses a direct match of the normalised waveform (Manhatten distance). Version 1.1 introduced as distance metric the correlation between normalized units. Concatenation quality is not yet included in the selection.

3.5. Soundscapes and Texture Resynthesis (2001)

The *Soundscapes* project [6] generates endless but never repeating soundscapes from a recording for installations. This means keeping the "texture" of the original sound file, while being able to play it for an arbitrarily long time. The segmentation into synthesis units is performed by a Wavelet analysis for good join points. This generative approach means that also the synthesis target is generated on the fly, driven by the original structure of the recording.

3.6. MoSievius (2003)

The MoSievius system [9] is an encouraging first attempt to apply unit selection to real-time performance-oriented synthesis with direct intuitive control. The system is based on units placed in a loop: A unit is played when its descriptor values lie within ranges controlled by the user. The feature set used contains voicing, energy, spectral flux, spectral centroid, instrument class. This method of content-based retrieval is called *Sound Sieve*.

3.7. Audio Mosaics (2003)

Audio mosaics [11], called “creative abuse” of MPEG-7 by their authors, calculated by finding the best matching snippets of one Beatles song, to reconstitute another one. The match was calculated from the MPEG-7 low-level descriptors, but no measure of concatenation quality was included in the selection.

3.8. Sound Clustering Synthesis (2003)

Kobayashi [8] resynthesises a target sound from a pre-analysed and pre-clustered sound base using a spectral match function. Resynthesis is done FFT-frame-wise, conserving the association of consecutive frame clusters. This leads to a good approximation of the synthesised sound with the target, and a high consistency in the development of the synthesised sound. Note that this does not necessarily mean a high spectral continuity, since also transitions from a release to an attack frame are captured by the pairwise association of database frame clusters.

3.9. Directed Soundtrack Synthesis (2003)

Audio and user directed sound synthesis [4] aims at the production of film soundtracks by replacing an existing one with sounds from a different audio source in small chunks similar in sound texture. It introduces user-definable constraints in the form of large-scale properties of the sound texture. For the unconstrained parts of the synthesis, a Hidden Markov Model based on the statistics of transition probabilities between spectrally similar sound segments is left running in generative mode, much similar to the approach of [6] described in section 3.5.

3.10. Let them sing it for you (2003)

A fun application of not-quite-CSS is this web site [1] , where a text given by a user is synthesised by looking up each word in a hand constituted monorepresented database of snippets of pop songs where that word is sung. The database is extended by user’s request for a new word. At the time of writing, it counted 1400 units.

3.11. Input Driven Resynthesis (2004)

This project [18] starts from a database of FFT frames from one week of radio recordings, analysed for loudness and 10 spectral bands as descriptors. The database then form a trajectories through the descriptor space. Phase vocoder resynthesis is controlled by live audio input that is analysed for the same descriptors, and the selection algorithm tries to follow a part of a database’s trajectory whenever possible, limiting jumps.

3.12. MATConcat (2004)

The *MATConcat* system [24] is an open source application in *Matlab* to explore concatenative resynthesis. For the moment, units are homogeneous windows taken out of the database sounds, so that this system is closer to controlled granular synthesis. The descriptors used are pitch, loudness, zero crossing rate, spectral centroid, spectral drop-off, and harmonicity, and selection is a match of descriptor values within a certain range of the target. Through the use of a large window function on the grains, the result sounds pleasingly smooth, which amounts to the squaring of the circle for concatenative synthesis. *MATConcat* is the first system used to compose two electroacoustic musical works, premiered at ICMC 2004: *Gates of Heaven and Hell* (concatenative variations on Mahler), and *Dedication to George Crumb*.

3.13. Synful (2004)

The first commercial application using some ideas of CSS is the *Synful* software synthesiser [2] [10], which aims at the reconstitution of expressive solo instrument performances from Midi input. Real instrument recordings are segmented into a database of attack, sustain, release, and transition units of varying subtypes. The real-time Midi input is converted by rules to a synthesis target that is then satisfied by selecting the closest units according to a simple pitch and loudness distance function. Synthesis is heavily using transformation of pitch, loudness, and duration, favoured by the hybrid waveform, spectral, and sinusoidal representation of the database units. *Synful* is more on the side of a rule-based sampler than CSS, with its fixed inventory and limited feature set, but fulfills the application of high-level instrument synthesis impressively well.

4. TAXONOMY

As a summary, we can order the above methods for concatenative musical synthesis according to two aspects, which combined indicate the level of “data-drivenness” of a method. They form the axes in figure 1, the abscissa indicating the structuredness of information obtained by analysis of the source sounds, and the ordinate the degree of automation of the selection. Further aspects are the inclusion of concatenation quality in the selection, and real-time capabilities.

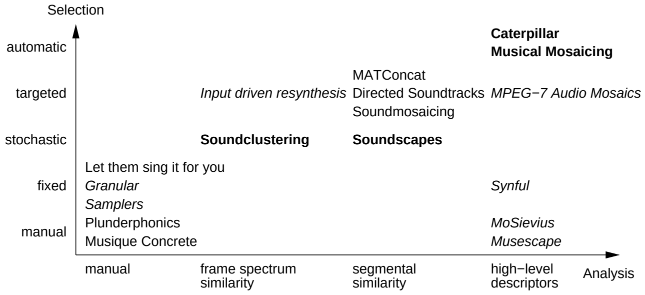

Figure 1. Comparison of musical sound synthesis methods according to selection and analysis, use of concatenation quality (bold), and real-time capabilities (italics)

Certain groups emanate clearly from this diagram:

Selection by hand with completely subjective manual analysis (*Musique Concrète*, *Plunderphonics*), with given tempo and character analysis (*drum loops*), or by a simple mapping (*Let them sing it for you*)

Selection by fixed mapping from a fixed inventory with some analysis in class and pitch (*samplers*), or a more flexible rule-based mapping together with a subset of automatic selection (*Synful*)

[1] http://www.sr.se/cgi-bin/p1/src/sing/default.asp

[2] http://www.synful.com

Arbitrary source, manual browsing, no analysis (*granular synthesis*)

Frame spectrum similarity analysis, target match selection (*Soundmosaicing*) with a partially stochastic selection (*Soundclustering*)

Segmental similarity analysis with stochastic (*Soundscapes*) or targeted selection (*Directed Soundtracks*, *MATConcat*, *Input Driven Resynthesis*)

Descriptor analysis with manual selection in real time (*MoSievius*, *Musescape*), or with fully automatic high-level unit selection and concatenation (*Caterpillar*, *Musical Mosaicing*) or without (*MPEG-7 audio mosaics*)

5. REMAINING PROBLEMS

Future work (described in more detail in [22]) could concentrate on the *segmentation* into units closer to the *sound object* as tempted in [6, 4, 7], or avoid a fixed segmentation altogether. Better *descriptors* augment the usability of the selection, e.g. for *percussiveness* [26], or by automatic discovery [28]. When the corpus is made of recordings of written music, *musical descriptors* can be obtained from an analysis of the score. *Mining the database* could provide data-driven target and concatenation distance functions. Meanwhile, the weights of the target distance can be optimised by exhaustive search in the weight-space as in [13], which also removes redundancies in the descriptors. *Real-time interactive selection* allows to browse a sound database. It needs a good model for navigation and efficient search algorithms.

6. CONCLUSION

We tried to show in this article that many approaches pick up the general idea of data-driven concatenative synthesis, or part of it, to achieve interesting results, without knowing about the other work in the field. To help the exchange of ideas and experience, a mailinglist *concat@ircam.fr* has been created, accessible from [23].

Concatenative synthesis from existing song material evokes tough legal questions of intellectual property, sampling and citation practices [15, 24]. Therefore, this ICMC's FreeSound project [3] is a welcome initiative.

Professional and multi-media sound synthesis shows a natural drive to make use of the advanced mass storage capacities available today, and the easily available large amount of digital content. We can foresee this type of applications hitting a natural limit of manageability of the amount of data. Only automatic support of the data-driven composition process will be able to surpass this limit and make the whole wealth of musical material accessible to the musician.

7. REFERENCES

[1] J.-J. Aucouturier, F. Pachet, and P. Hanappe. From sound sampling to song sampling. In *International Symposium on Music Information Retrieval (ISMIR)*, 2004.

[2] G. Beller. Un synthétiseur vocal par sélection d'unités. Master's thesis, Ircam, Paris, 2004.

[3] G. Beller, D. Schwarz, T. Hueber, and X. Rodet. A hybrid concatenative synthesis system on the intersection of music and speech. In *Journées d'Informatique Musicale (JIM)*, St. Denis, 2005.

[4] M. Cardle, S. Brooks, and P. Robinson. Audio and user directed sound synthesis. In *Proc. ICMC*, Singapore, 2003.

[5] Steven Hazel. Soundmosaic. web page, 2001. http://thalassocracy.org/soundmosaic.

[6] R. Hoskinson and D. Pai. Manipulation and resynthesis with natural grains. In *Proc. ICMC*, Havana, 2001.

[7] T. Jehan. Event-synchronous music analysis/synthesis. In *Digital Audio Effects (DAFx)*, Naples, 2004.

[8] R. Kobayashi. Sound clustering synthesis using spectral data. In *Proc. ICMC*, Singapore, 2003.

[9] Ari Lazier and Perry Cook. MOSIEVIUS: Feature driven interactive audio mosaicing. In *Digital Audio Effects (DAFx)*, London, 2003.

[10] E. Lindemann. Musical synthesizer capable of expressive phrasing. US Patent 6,316,710, 2001.

[11] A. T. Lindsay and M. Casey. Sound Replacement, Beat Unmixing and Audio Mosaics: Content-Based Audio Processing with MPEG-7. Digital Audio Effects (DAFx) Workshop London, 2003.

[12] Adam T. Lindsay, Alan P. Parkes, and Rosemary A. Fitzgerald. Description-driven context-sensitive effects. In *Digital Audio Effects (DAFx)*, London, 2003.

[13] Yoram Meron. *High Quality Singing Synthesis Using the Selection-based Synthesis Scheme*. PhD thesis, University of Tokyo, 1999.

[14] N. Orio and D. Schwarz. Alignment of monophonic and polyphonic music to a score. In *ICMC*, Havana, 2001.

[15] John Oswald. Plunderphonics. web page, 1999. http://www.plunderphonics.com.

[16] F. Pachet, P. Roy, and D. Cazaly. A combinatorial approach to content-based music selection. *IEEE MultiMedia*, 7(1), 2000.

[17] G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of MPEG-7. In *Proc. ICMC*, Berlin, 2000.

[18] M. Puckette. Low-dimensional parameter mapping using spectral envelopes. In *Proc. ICMC*, Miami, 2004.

[19] D. Schwarz. A system for data-driven concatenative sound synthesis. In *Digital Audio Effects (DAFx)*, Verona, 2000.

[20] D. Schwarz. The CATERPILLAR system for data-driven concatenative sound synthesis. In *Digital Audio Effects (DAFx)*, London, 2003.

[21] D. Schwarz. New developments in data-driven concatenative sound synthesis. In *Proc. ICMC*, Singapore, 2003.

[22] D. Schwarz. *Data-Driven Concatenative Sound Synthesis*. PhD thesis, Université Paris 6, 2004.

[23] D. Schwarz. Caterpillar. Web page, 2005. http://recherche.ircam.fr/anasyn/schwarz/thesis

[24] Bob L. Sturm. MATConcat: An application for exploring concatenative sound synthesis using Matlab. In *Proc. ICMC*, Miami, 2004.

[25] G. Tzanetakis. MUSESCAPE: An interactive content-aware music browser. In *Digital Audio Effects (DAFx)*, London, 2003.

[26] G. Tzanetakis, G. Essl, and P. Cook. Human perception and computer extraction of musical beat strength. In *Digital Audio Effects (DAFx)*, Hamburg, 2002.

[27] A. Zils and F. Pachet. Musical mosaicing. In *Digital Audio Effects (DAFx)*, Limerick, 2001.

[28] A. Zils and F. Pachet. Extracting automatically the perceived intensity of music titles. In *Digital Audio Effects (DAFx)*, London, 2003.

[3] http://iua-freesound.upf.es/