CONCATENATIVE SOUND SYNTHESIS: THE EARLY YEARS

Diemo Schwarz
Ircam – Centre Pompidou
1, place Igor-Stravinsky, 75003 Paris, France
http://www.ircam.fr/anasyn/schwarz http://concatenative.net
schwarz@ircam.fr

ABSTRACT

Concatenative sound synthesis is a promising method of musical sound synthesis with a steady stream of work and publications for over five years now. This article offers a comparative survey and taxonomy of the many different approaches to concatenative synthesis throughout the history of electronic music, starting in the 1950s, even if they weren't known as such at their time, up to the recent surge of contemporary methods. Concatenative sound synthesis methods use a large database of source sounds, segmented into *units*, and a *unit selection* algorithm that finds the units that match best the sound or musical phrase to be synthesised, called the *target*. The selection is performed according to the descriptors of the units. These are characteristics extracted from the source sounds, e.g. pitch, or attributed to them, e.g. instrument class. The selected units are then transformed to fully match the target specification, and 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 most urgent and interesting problems for further work on concatenative synthesis are listed concerning segmentation, descriptors, efficiency, legality, data mining, and real time interaction. Finally, the conclusion tries to provide some insight into the current and future state of concatenative synthesis research. [1]

1. INTRODUCTION

When technology advances and is easily accessible, creation progresses, too, driven by the new possibilities that are open to be explored. For musical creation, we have seen such surges of creativity throughout history, for example with the first easily usable recording devices in the 1940s, with widespread diffusion of electronic synthesizers from the 1970s, and with the availability of real-time interactive digital processing tools at the end of the 1990s.

The next relevant technology advance is already here, widespread diffusion just around the corner, and waiting to be exploited for creative use: Large databases of sound, with a pertinent description of their contents, ready for content-based retrieval. These databases want to be exploited for musical sound synthesis, and concatenative synthesis looks like the natural candidate to do so.

Concatenative sound synthesis (CSS) 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, which always degrade sound quality, is reduced. The units can be *non-uniform* (heterogeneous), i.e. they can comprise a sound snippet, an instrument note, up to a whole phrase. Most often, however, a homogeneous size and type of units is used, and sometimes a unit is just a short time window of the signal used in conjunction with spectral analysis and overlap-add synthesis.

Usual sound synthesis methods are based on a model of the sound signal. It is very difficult to build a model that would realistically generate all the fine details of the sound. Concatenative synthesis, on the contrary, by using actual recordings, preserves entirely these details. For example, very naturally sounding transitions can be synthesized, since unit selection is aware of the context of the database units. In this *data-driven approach*, instead of supplying rules constructed by careful thinking as in a *rule-based approach*, the rules are induced from the data itself. Findings in other domains, e.g. speech recognition, corroborate the general superiority of data-driven approaches. Concatenative synthesis can be more or less data-driven; more is advantageous because the information contained in the many sound examples in the database can be exploited. This will be the main criterion for the taxonomy of approaches to concatenative synthesis in section 3.

Concatenative synthesis sprung up independently in multiple places and 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. In this article, we try to acknowledge this young field of musical sound synthesis that has been identified as such only five years ago. Many fields and topics of research intervene, examples of which are given in section 2.

Development has accelerated over the past few years as can be seen in the presentation and comparison of the different approaches and systems in section 3. There are

[1] This is a preprint of an article whose final and definitive form has been published in the Journal of New Music Research vol. 35 num. 1, March 2006 [copyright Taylor & Francis]. Journal of New Music Research is available online at: http://journalsonline.tandf.co.uk

now the first commercial products available (3.2.4, 3.2.5), and, last but not least, ICMC 2004 saw the first musical pieces using concatenative synthesis (3.4.5).

Section 4 finally gives some of the most urgent problems to be tackled for the further development of concatenative synthesis.

1.1. Applications

The current work on concatenative synthesis focuses on four 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. This hypothesis is illustrated in figure 1.

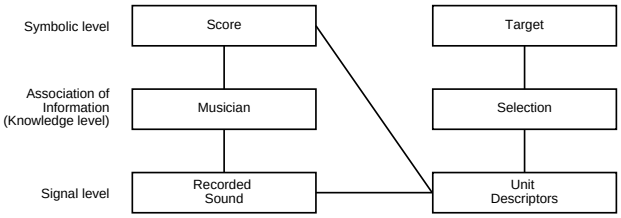


Figure 1. Hypothesis of high level synthesis: The relations between the score and the produced sound in the case of performing an instrument, and the synthesis target and the unit descriptors in the case of concatenative data-driven synthesis are shown on their respective level of representation of musical information. [3]

Resynthesis of audio with sounds from the database: A sound or phrase is taken as the audio score, which is resynthesized with the sequence of units best matching its descriptors, e.g., with the same pitch, amplitude, and/or timbre characteristics.

This is often referred to as *audio mosaicing*, since it tries to reconstitute a given larger entity from many small parts as in the recently popular *photo mosaics*.

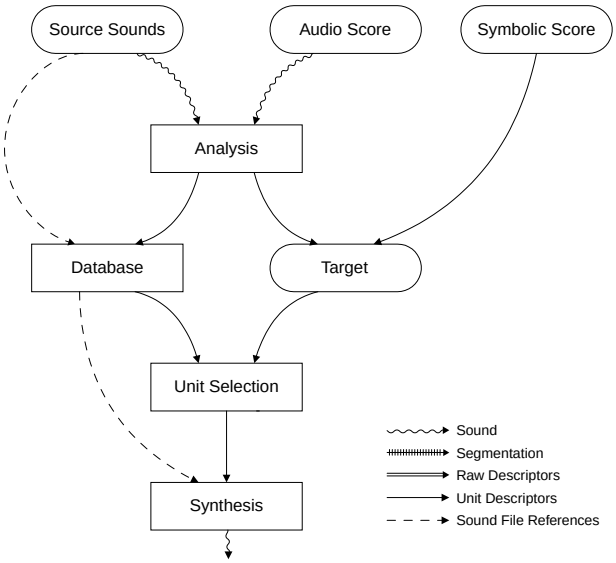
Texture and ambience synthesis is used for installations or film production. It aims at generating soundtracks from sound libraries or preexisting ambience recordings, or extending soundscape recordings for an arbitrarily long time, regenerating the character and flow but at the same time being able to control larger scale parameters.

Free synthesis from heterogeneous sound databases offers a sound composer efficient control of the result by using perceptually meaningful descriptors to specify a target as a multi-dimensional curve in the descriptor space. If the selection happens in real-time, this allows to browse and explore a corpus of sounds interactively.

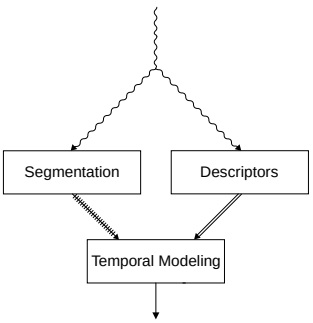
[3] According to Vinet (2003), we can classify digital musical representations into the physical level, the signal level, the symbolic level, and the knowledge level.

1.2. Technical Overview

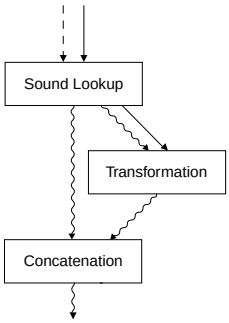
Any concatenative synthesis system performs the tasks illustrated in figure 2, sometimes implicitly. This list of tasks will serve later for our taxonomy of systems in section 3.



(a) General structure.



(b) Analysis component.



(c) Synthesis component.

Figure 2. Data flow model of a concatenative synthesis system, rounded boxes representing data, rectangular boxes components, and arrows flow of data.

1.2.1. 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 segmentation according to transients or spectral change, or arbitrary grain segmentation for free and re-synthesis, or can happen on-the-fly.

The descriptors can be of type categorical (a class membership), static (a constant text or numerical value for a unit), or dynamic (varying over the duration of a unit), and from one of the following classes: category (e.g. instrument), signal, symbolic, score, perceptual, spectral, harmonic, or segment descriptors (the latter serve for bookkeeping). Descriptors are usually analysed by au-

tomatic methods, but can also be given as external metadata, or be supplied by the user, e.g. for categorical descriptors or for subjective perceptual descriptors. (e.g. a "glassiness" value or "anxiousness" level could be manually attributed to units).

For the time-varying dynamic descriptors, temporal modeling reduces the evolution of the descriptor value over the unit to a fixed-size vector of values characterizing this evolution. Usually, only the mean value is used, but some systems go further and store range, slope, min, max, attack, release, modulation, and spectrum of the descriptor curve.

1.2.2. Database

Source file references, units, unit descriptors, and the relationships between them are stored in a database. The subset of the database that is preselected for one particular synthesis is called the *corpus*. Often, the database is implicitly constituted by a collection of files. More rarely, a (relational or other) database management system is used, which can run locally or on a server. Internet sound databases with direct access to sounds and descriptors [4] are beginning to make their appearance, e.g. with the *freesound* project (see section 4.3).

1.2.3. Target

The target is specified as a sequence of target units with their desired descriptor characteristics. Usually, only a subset of the available database descriptors is given. The unspecified descriptors do not influence the selection directly, but can, however, be used to stipulate continuity via the concatenation distance (see section 1.2.6 below). The target can either be generated from a symbolic score (expressed e.g. in notes or directly in segments plus descriptors), or analysed from an audio score (using the same segmentation and analysis methods as for the source sounds).

1.2.4. Selection

The unit selection algorithm is crucial as it contains all the "intelligence" of data-driven concatenative synthesis. Units are selected from the database that match best the given sequence of target units and descriptors according to a distance function (section 1.2.5) and a concatenation quality function (section 1.2.6). The selection can be local (the best match for each target unit is found individually), global (the sequence with the least total distance is found), or iterative (by a search algorithm that approaches the globally optimal selection until a maximum number of search steps is reached).

Two different classes of algorithms can be found in the approaches described in this article: path-search unit selection (section 1.2.7), and unit selection based on a constraint solving approach (section 1.2.8).

Most often, however, a simple local search for the best matching unit is used without taking care of the context. In some real-time approaches, the local context between the last selected unit and all matching candidates for the following unit is considered. Both local possibilities can be seen as a simplified form of the path search unit selection algorithm, which still uses the same framework of distance functions, presented in its most general formulation in the following

1.2.5. Target Distance

The target distance C^t corresponds to the perceptual similarity of the database unit u_i to the target unit t_τ. It is given as a sum of p weighted individual descriptor distance functions C_k^t as:

$$C^t(u_i, t_\tau) = \sum_{k=1}^{p} w_k^t \, C_k^t(u_i, t_\tau) \tag{1}$$

To favour the selection of units out of the same context in the database as in the target, the *context distance* C^x considers a sliding context in a range of r units around the current unit with weights w_j decreasing with distance j.

$$C^x(u_i, t_\tau) = \sum_{j=-r}^{r} w_j^x \, C^t(u^{i+j}, t^{\tau+j}) \tag{2}$$

Mostly, a Euclidean distance normalised by the standard deviation is used and r is zero. Some descriptors need specialised distance functions. Symbolic descriptors, e.g. phoneme class, require a lookup table of distances.

1.2.6. Concatenation Distance

The concatenation distance C^c expresses the discontinuity introduced by concatenating the units u_i and u_j from the database. It is given by a weighted sum of q descriptor concatenation distance functions C_k^c:

$$C^c(u_i, u_j) = \sum_{k=1}^{q} w_k^c \, C_k^c \, (u_i, u_j) \tag{3}$$

The distance depends on the unit type: concatenating an attack unit allows discontinuities in pitch and energy, a sustain unit does not. Consecutive units in the database have a concatenation distance of zero. Thus, if a whole phrase matching the target is present in the database, it will be selected in its entirety.

1.2.7. The Path Search Unit Selection Algorithm

This unit selection algorithm is based on the standard path search algorithm used in speech synthesis, first proposed by Hunt and Black (1996). It has been adapted to the specificities of musical sound synthesis for the first time by Schwarz (2000) in the *Caterpillar* system described in section 3.7.1.

The unit database can be seen as a fully connected state transition network through which the unit selection algorithm has to find the least costly path that constitutes the target. Using the weighted extended target distance $w^t C^x$ as the *state occupancy cost*, and the weighted concatenation distance $w^c C^c$ as the *transition cost*, the optimal path can be efficiently found by a Viterbi algorithm (Viterbi, 1967; Forney, 1973). A detailed formulation of the algorithm is given by Schwarz (2004).

[4] This excludes the many existing web collections of sounds accessed by a search term found in the title, e.g. http://sound-effects-library.com.

1.2.8. Unit Selection by Constraint Solving

Applying the formalism of *constraint satisfaction* to unit selection permits to express musical desiderata additional to the target match in a flexible way, such as to avoid repeating units, or not to use a certain unit for the selection. It has been first proposed for music program generation by Pachet, Roy, and Cazaly (2000), see section 2.4, and for data-driven concatenative musical synthesis by Zils and Pachet (2001) in the *Musical Mosaicing* system described in section 3.7.3.

It is based on the *adaptive local search* algorithm described in detail in (Codognet & Diaz, 2001; Truchet, Assayag, & Codognet, 2001), which runs iteratively until a satisfactory result is achieved or a certain number of iterations is reached. Constraints are here given by an error function, which allows us to easily express the unit selection algorithm as a constraint satisfaction problem (CSP) using the target and concatenation distances between units.

1.2.9. Synthesis

The final waveform synthesis is done by concatenation of selected units with a short cross-fade, possibly applying transformations, for instance altering pitch or loudness. 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, speech or texture synthesis).

2. RELATED TOPICS

Concatenative synthesis is at the intersection of many fields of research, such as music information retrieval (MIR), database technology, real-time and interactive methods, digital signal processing (DSP), sound synthesis models, musical modeling, classification, perception.

We could see concatenative synthesis as one of three variants of content-based retrieval, depending on what is queried and how it is used. When just one sound is queried, we are in the realm of descriptor- or similarity-based sound selection. Superposing retrieved sounds to satisfy a certain outcome is the topic of automatic orchestration tools (Hummel, 2005). Finally, sequencing retrieved sound snippets is our topic of concatenative synthesis.

Other closely related research topics are given in the following, that share many of the basic questions and problems.

2.1. Speech Synthesis

Research in musical synthesis is heavily influenced by research in speech synthesis, which can be said to be roughly 10 years ahead. Concatenative unit selection speech synthesis from large databases (Hunt & Black, 1996) is used in a great number of Text-to-Speech systems for waveform generation (Prudon, 2003). Its introduction resulted in a considerable gain in quality of the synthesized speech over rule-based parametric synthesis systems in terms of naturalness and intelligibility. Unit selection algorithms attempt to estimate the appropriateness of a particular database speech unit using linguistic features predicted from a given text to be synthesized. The units can be of any length (non-uniform unit selection), from sub-phonemes to whole phrases, and are not limited to diphones or triphones.

Although concatenative sound synthesis is quite similar to concatenative speech synthesis and shares many concepts and methods, both have different goals. Even from a very rough comparison between musical and speech synthesis, some profound differences spring to mind, which make the application of concatenative data-driven synthesis techniques from speech to music non-trivial:

- Speech is a-priori clustered into phonemes. A musical analogue for this *phonemic identity* are pitch classes which are applicable for tonal music, but in general, no a-priori clustering can be presupposed.
- In speech, the time position of synthesized units is intrinsically given by the required duration of the selected units. In music, precise time-points have to be hit when we want to keep the rhythm.
- In speech synthesis, intelligibility and naturalness are of prime interest, and the synthesised speech is often limited to "normal" informative mode. However, musical creation is based on artistic principles, uses many modes of expressivity, and needs to experiment. Therefore, creative and interactive use of the system should be possible by using any database of sounds, any descriptors, and a flexible expression of the target for the selection.

2.2. Singing Voice Synthesis

Concatenative singing voice synthesis occupies an intermediate position between speech and sound synthesis, whereas the used methods are most often closer to speech synthesis, [5] with the limitation of fixed inventories specifically recorded, such as the *Lyricos* system (Macon et al., 1997a, 1997b), the work by Lomax (1996), and the recent system developed by Bonada et al. (2001). There is one notable exception (Meron, 1999), where an automatically constituted large unit database is used.

See (Rodet, 2002) for an up-to-date overview of current research in singing voice synthesis, which is out of the scope of this article.

This recent spread of data-driven singing voice synthesis methods based on unit selection follows their success in speech synthesis, and lets us anticipate a coming leap in quality and naturalness of the singing voice. Regarding the argument of rule-based vs. data-driven singing voice synthesis, Rodet (2002) notes that:

> *Clearly, the units intrinsically contain the influence of an implicit set of rules applied by the*

[5] Concatenative speech synthesis techniques are directly used for singing voice synthesis in *Burcas* (http://www.ling.lu.se/persons/Marcusu/music/burcas), *Flinger* (http://www.cslu.ogi.edu/tts/flinger), and abused in http://www.silexcreations.com/melissa.

singer with all his training, talent and musical skill. The unit selection and concatenation method is thus a way to replace a large and complicated set of rules by implicit rules from the best performers, and it is often called a data-driven concatenative synthesis.

2.3. Content-Based Processing

Content-based processing is a new paradigm in digital audio processing that is based on symbolic or high-level manipulations of elements of a sound, rather than using signal processing alone (Amatriain et al., 2003). Lindsay, Parkes, and Fitzgerald (2003) propose context-sensitive effects that are more aware of the structure of the sound than current systems by utilising content descriptions such as those enabled by MPEG-7 (Thom, Purnhagen, Pfeiffer, & MPEG Audio Subgroup, 1999; Hunter, 1999). Jehan (2004) works on object-segmentation and perception-based description of audio material and then performs manipulations of the audio in terms of its musical structure. The *Song Sampler* (Aucouturier, Pachet, & Hanappe, 2004) is a system which automatically samples parts of a song, assigns it to the keys of a MIDI-keyboard to be played with by a user.

2.4. Music Selection

The larger problem of music selection from a catalog has some related aspects with selection-based sound synthesis. Here, the user wants to select a sequence of songs (a compilation or *playlist*) according to his taste and a desired evolution of high-level features from one song to the next, e.g. augmenting tempo and percieved energy. The problem is well described in (Pachet et al., 2000), and an innovative solution based on constraint satisfaction is proposed, which ultimately inspired the use of constraints for sound synthesis in (Zils & Pachet, 2001), see section 3.7.3.

Other music retrieval systems approach the problematic of selection: The *Musescape* music browser (Tzanetakis, 2003) works with an intuitive and space-saving interface by specifying high-level musical descriptors (tempo, genre, year) on sliders. The system then selects in real time musical excerpts that match the desired descriptors.

3. TAXONOMY

Approaches to musical sound synthesis that are somehow data-driven and concatenative can be found throughout history. The earlier uses are usually not identified as such, but the brief discussion in this section argues that they can be seen as instances of fixed inventory or manual concatenative synthesis. I hope to show that all these approaches are very closely related to, or can sometimes even be seen as a special case of the general formulation of concatenative synthesis in section 1.2.

Table 1 lists in chronological order all the methods for concatenative musical sound synthesis that will be discussed in the following, proposing several properties for comparison. We can order these methods according to two main aspects, which combined indicate the level of "data-drivenness" of a method. They form the axes of the diagram in figure 3, the abscissa indicating the degree of structuredness of information obtained by analysis of the source sounds and the metadata, and the ordinate the degree of automation of the selection. Further aspects expressed in the diagram are the inclusion of concatenation quality in the selection, and real-time capabilities.

Groups of similar approaches emanate clearly from this diagram that will be discussed in the following seven sub-sections, going from left to right and bottom to top through the diagram.

3.1. Group 1: Manual Approaches

These historical approaches to musical composition use selection by hand with completely subjective manual analysis (*Musique Concrète*, *Plunderphonics*) or based on given tempo and character analysis (*phrase sampling*). It is to note that these approaches are the only ones described here that aim, besides sequencing, also at layering the selected sounds.

For musical sound synthesis (leaving aside the existing attempts for *collage* type sonic creations), we'll start by shedding a little light on some preceding synthesis techniques, starting from the very beginning when recorded sound became available for manipulation:

3.1.1. Musique Concrète and Early Electronic Music (1948)

Going very far back, and extending the term far beyond reason, "concatenative" synthesis started with the invention of the first usable recording devices in the 1940's: the phonograph and, from 1950, the magnetic tape recorder (Battier, 2001, 2003). The tape cutting and splicing techniques were advanced to a point that different types of diagonal cuts were applied to control the character of the concatenation (from an abrupt transition to a more or less smooth cross-fade).

3.1.2. Pierre Schaeffer

The *Groupe de Recherche Musicale* (GRM) of Pierre Schaeffer used for the first time recorded segments of sound to create their pieces of *Musique Concrète*. In the seminal work *Traité des Objets Musicaux* (Schaeffer, 1966), explained in (Chion, 1995), Schaeffer defines the notion of *sound object*, which is not so far from what is here called *unit*: A sound object is a clearly delimited segment in a source recording, and is the basic unit of composition (Schaeffer & Reibel, 1967, 1998). Moreover, Schaeffer strove to base his theory of sound analysis on objectively, albeit manually, observable characteristics, the *écoute réduite* (narrow listening) (GRAM, 1996), which corresponds to a standardised descriptor set of the perceptible qualities of mass, grain, duration, matter, volume, and so on.

Group, Name (Author)	Year	Type	Application	Inventory	Units	Segmentation	Descriptors	Selection	Concatenation	Real-time
1 Musique Concrete (Schaeffer)	1948	art	composition	open	heterogeneous	manual	manual	manual	manual	no
2 Digital Sampling	1980	sound	high-level	fixed	notes/any	manual	manual	fixed mapping	no	yes
1 Phrase Sampling	1990	art	composition	open	phrases	manual	musical	manual	no	no
2 Granular Synthesis	1990	sound	free	open	homogeneous	fixed	time	manual	no	yes
1 Plunderphonics (Oswald)	1993	art	composition	open	heterogeneous	manual	manual	manual	manual	no
7 Caterpillar (Schwarz)	2000	research	high-level	open	heterogeneous	alignment	high-level	global	yes	no
7 Musaicing (Pachet et al.)	2001	research	resynthesis	open	homogeneous	blind	low-level	constraints	no	no
4 Soundmosaic (Hazel)	2001	application	resynthesis	open	homogeneous	fixed	signal match	local	no	no
4 Soundscapes (Hoskinson et al.)	2001	application	texture	open	homogeneous	automatic	signal match	local	yes	no
3 La Légende des siècles (Pasquet)	2002	sound	resynthesis	open	frames	blind	spectral match	spectral	no	yes
4 Granuloop (Xiang)	2002	rhythm	free	open	beats	beat	spectral match	local	yes	yes
5 MoSievius (Lazier and Cook)	2003	research	free	open	homogeneous	blind	low-level	local	no	yes
5 Musescape (Tzanetakis)	2003	research	music selection	open	homogeneous	blind	high-level	local	no	yes
6 MPEG-7 Audio Mosaics (Casey and Lindsay)	2003	research	resynthesis	open	homogeneous	on-the-fly	low-level	local	no	yes
3 Sound Clustering Synthesis (Kobayashi)	2003	research	resynthesis	open	frames	fixed	low-level	spectral	no	no
4 Directed Soundtrack Synthesis (Cardle et al.)	2003	research	texture	open	heterogeneous	automatic	low-level	constraints	yes	no
2 Let them sing it for you (Bunger)	2003	web art	high-level	fixed	words	manual	semantic	direct	no	no
6 Network Auralization for Gnutella (Freeman)	2003	software art	high-level	open	homogeneous	blind	context-dependent	local	no	yes
3 Input driven resynthesis (Puckette)	2004	research	resynthesis	open	frames	fixed	low-level	local	yes	yes
4 Matconcat (Sturm)	2004	research	resynthesis	open	homogeneous	fixed	low-level	local	no	no
2 Synful (Lindemann)	2004	commercial	high-level	fixed	note parts	manual	high-level	lookahead	yes	yes
6 SoundSpotter (Casey)	2005	research	resynthesis	open	homogeneous	on-the-fly	morphological	local	no	yes
7 Audio Analogies (Simon et al.)	2005	research	high-level	open	notes/dinotes	manual	pitch	global	yes	no
7 Ringomatic (Aucouturier et al.)	2005	research	high-level	open	drum bars	automatic	high-level	global	yes	yes
5 frelia (Momeni and Mandel)	2005	installation	free	open	homogeneous	none	high-level+abstract	local	no	yes
5 CataRT (Schwarz)	2005	sound	free	open	heterogeneous	alignment/blind	high-level	local	no	yes
2 Vienna Symphonic Library Instruments	2006	commercial	high-level	fixed	note parts	manual	high-level	lookahead	yes	yes
6 iTunes Signature Maker (Freeman)	2006	software art	high-level	open	homogeneous	blind	context-dependent	local	no	yes

Table 1. Comparison of concatenative synthesis work in chronological order

3.1.3. Karlheinz Stockhausen

Schaeffer (1966) also relates Karlheinz Stockhausen's desire to cut a tape into millimeter-sized pieces to recompose them, the notorious *Étude des 1000 collants* (study with one thousand pieces) of 1952. The piece (actually simply called *Étude*) was composed according to a score generated by a series for pitch, duration, dynamics, and timbral content, for a corpus of recordings of hammered piano strings, transposed and cropped to their steady sustained part (Manion, 1992).

3.1.4. John Cage

John Cage's *Williams Mix* (1953) is a composition for 8 magnetic tapes that prescribes a corpus of about 600 recordings in 6 categories (e.g. city sounds, country sounds, electronic sounds), and how they are to be ordered and spliced together (Cage, 1962). [6] [7]

3.1.5. Iannis Xenakis

In Iannis Xenakis' *Analogique A et B* (1958/1959) the electronic part *B* is composed of cut and spliced pieces of tape, selected according to a stochastic process. The orchestral part *A* is supposed to be an analogue to *B*, where these "units" are realised by acoustic instruments. They are here expressed as half bar pieces of a score, stochastically selected from an (implicit) corpus according to *pitch group*, *dynamics*, and *density* (DiScipio, 2005).

[6] http://www.medienkunstnetz.de/works/williams-mix
[7] http://www.johncage.info/workscage/williamsmix.html

3.1.6. Phrase Sampling (1990's)

In commercial, mostly electronic, dance music, a large part of the musical material comes from specially constituted sampling CDs, containing rhythmic loops and short bass or melodic phrases. These phrases are generally labeled and grouped by tempo and sometimes characterised by mood or atmosphere. As the available CDs, aimed at professional music producers, number in the tens of thousands, each containing hundreds of samples, a large part of the work still consists in listening to the CDs and selecting suitable material that is then placed on a rhythmic grid, effectively constituting the base of a new song by concatenation of preexisting musical phrases.

3.1.7. Plunderphonics (1993)

Plunderphonics (Oswald, 1999) is John Oswald's artistic project of cutting up recorded music. One outstanding example, *Plexure*, is made up from 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, which were the descriptors used to guide the selection:

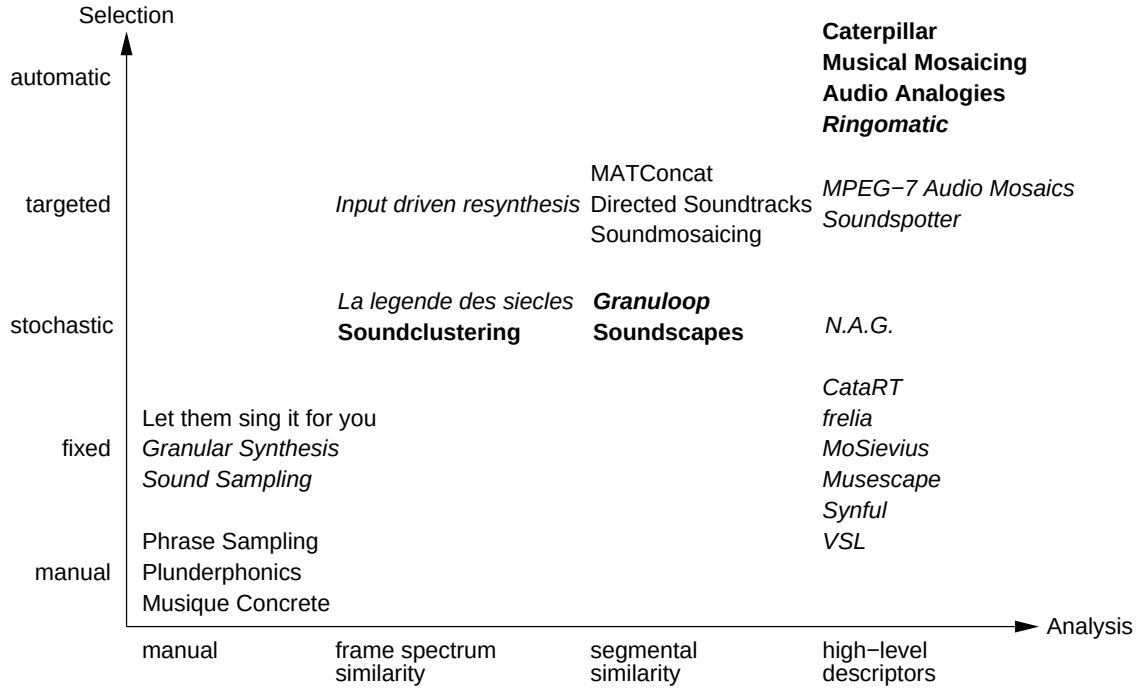


Figure 3. Comparison of musical sound synthesis methods according to selection and analysis, use of concatenation quality (bold), and real-time capabilities (italics)

> *Plundered are over a thousand pop stars from the past 10 years. [...] It starts with rapmillisyllables and progresses through the material according to tempo (which has an interesting relationship with genre).*
>
> Oswald (1993)

Cutler (1994) gives an extensive account of Oswald's and related work throughout art history and addresses the issue of the incapability of copyright laws to handle this form of musical composition.

3.2. Group 2: Fixed Mapping

Here, the selection is performed by a predetermined mapping from a fixed inventory with no analysis at all (*granular synthesis*), manual analysis (*Let them sing it for you*), some analysis in class and pitch (*digital sampling*), or a more flexible rule-based mapping that takes care of selecting the appropriate transitions from the last selected unit to the next in order to obtain a good concatenation (*Synful*, *Vienna Symphonic Library*).

3.2.1. Digital Sampling (1980's)

In the widest reasonable sense of the term, *digital sampling synthesisers* or *samplers* for short, which appeared at the beginning of the 1980's, were the first "concatenative" sound synthesis devices. A sampler is a device that can digitally record sounds and play them back, applying transposition, volume changes, and filters. Usually the recorded sound would be a note from an acoustic instrument, that is then mapped to the sampler's keyboard. *Multisampling* uses several notes of different pitches, and also played with different dynamics, to better capture the timbral variations of the acoustic instrument (Roads, 1996).

Modern software samplers can use several gigabytes of sound data [8] which makes samplers clearly a data-driven fixed-inventory synthesis system, with the sound database analysed by instrument class, playing style, pitch, and dynamics, and the selection being reduced to a fixed mapping of MIDI-note and velocity to a sample, without paying attention to the context of the notes played before, i.e. no consideration of concatenation quality.

3.2.2. Granular Synthesis (1990's)

Granular synthesis (Roads, 1988, 2001) takes short snippets out of a sound file called *grains*, 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 data-driven, but 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 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, as described in the free synthesis application in section 1.1.

[8] For instance, Nemesys, the makers of *Gigasampler*, [9] pride themselves to have sampled every note of a grand piano in every possible dynamic, resulting in a 1 GB sound set.

[9] http://www.nemesysmusic.com

3.2.3. Let them sing it for you (2003)

A fun web art project and application of not-quite-CSS is this site [10] (Bünger, 2003), 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 about 2000 units.

3.2.4. Synful (2004)

The first commercial application using some ideas of CSS is the the *Synful* software synthesiser [11] (Lindemann, 2001), based on the technology of *Reconstructive Phrase Modeling*, 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 of the appropriate type 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 set of descriptors, but fulfills the application of high-level instrument synthesis (section 1.1) impressively well.

3.2.5. Vienna Symphonic Library (2006)

The *Vienna Symphonic Library* [12] is a huge collection (550 GB) of samples of all classical instruments in all playing styles and moods, and including single notes, groups of notes, and transitions. Their so-called *performance detection* algorithms offer the possibility to automatically analyse a MIDI performance input and to select samples appropriate for the given transition and context in a real-time instrument plugin.

3.3. Group 3: Spectral Frame Similarity

This subclass of data-driven synthesis uses as units short-time signal frames that are matched to the target by a spectral similarity analysis (*Input Driven Resynthesis*) or additionally with a partially stochastic selection (*La Légende des siècles*, *Sound Clustering*). Here, forced by the short unit length, the selection must take care of the local context by stipulating certain continuity constraints, because otherwise FFT-frame salad would result.

3.3.1. La Légende des siècles (2002)

La Légende des siècles is a theatre piece performed at the *Comédie Française*, using real-time transformation on readings of Victor Hugo. One of these effects, developed by Olivier Pasquet, uses a data-driven synthesis method inspired by CSS: Prerecorded audio is analysed off-line frame-by-frame according to the descriptors energy and pitch. Each FFT frame is then stored in a dictionary and is clustered using the statistics program *R* [13] . During the performance, this dictionary of FFT-frames is used with an inverse FFT and overlap-add to resynthesize sound according to a target specification of pitch and energy. The continuity of the resynthesized frames is assured by a Hidden Markov Model trained on the succession of FFT-frame classes in the recordings.

3.3.2. Sound Clustering Synthesis (2003)

Kobayashi (2003) resynthesises a target given by a classical music piece from a pre-analysed and pre-clustered sound base using a vector-based direct spectral match function. Resynthesis is done FFT-frame-wise, conserving the association of consecutive frame clusters, i.e. the current frame to be synthesised will be similar to the current target frame, and the transition from one frame to the next will be similar to one occuring in the sound data, in the same context. 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 note release to an attack frame are captured by the pairwise association of database frame clusters.

3.3.3. Input Driven Resynthesis (2004)

This project (Puckette, 2004) starts from a database of FFT frames from one week of radio recording, analysed for loudness and 10 spectral bands as descriptors. The recording then forms a trajectory through the descriptor space mapped to a hypersphere. Phase vocoder overlap–add resynthesis is controlled in real-time by audio input that is analysed for the same descriptors, and the selection algorithm tries to follow a part of the database's trajectory whenever possible, limiting jumps.

3.4. Group 4: Segmental Similarity

This group's units are homogeneous segments that are locally selected by stochastic methods (*Soundscapes and Textures*, *Granuloop*), or matched to a target by segment similarity analysis on low-level signal processing descriptors (*Soundmosaicing*, *Directed Soundtracks*, *MATConcat*).

3.4.1. Soundscapes and Texture Resynthesis (2001)

The *Soundscapes* [14] project (Hoskinson & Pai, 2001) 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. A similar aim and approach is described in (Dubnov, Bar-Joseph, El-Yaniv, Lischinski, & Werman,

[10] http://www.sr.se/sing
[11] http://www.synful.com
[12] http://www.vsl.co.at
[13] http://www.r-project.org
[14] http://www.cs.ubc.ca/~reynald/applet/Scramble.html

2002). This generative approach means that also the synthesis target is generated on the fly, driven by the original structure of the recording.

3.4.2. Soundmosaic (2001)

Soundmosaic (Hazel, 2001) constructs an approximation of one sound out of small pieces of varying size from other sounds (called *tiles*). For version 1.0 of *Soundmosaic*, the selection of the best source tile uses a direct match of the normalised waveform (Manhatten distance). Version 1.1 introduced as distance metric the correlation between normalized tiles (the dot product of the vectors over the product of their magnitudes). Concatenation quality is not yet included in the selection.

3.4.3. Granuloop (2002)

The data-driven probabilistic drum loop rearranger *Granuloop* [15] (Xiang, 2002) is a patch for *Pure Data* [16] , which constructs transition probabilities between 16th notes from a corpus of four drum loops. These transitions then serve to exchange segments in order to create variation, either autonomously or with user interaction.

The transition probabilities (i.e. the concatenation distances) are analysed by loudness and spectral similarity computation, in order to favour continuity.

3.4.4. Directed Soundtrack Synthesis (2003)

Audio and user directed sound synthesis (Cardle, Brooks, & Robinson, 2003; Cardle, 2004) is aimed at the production of soundtracks in video by replacing existing soundtracks 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, e.g. preferred audio clips that shall appear at a certain moment. 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 freely in generative mode, much similar to the approach of Hoskinson and Pai (2001) described in section 3.4.1.

A slighly different approach is taken by Cano et al. (2004), where a sound atmosphere library is queried with a search term. The resulting sounds, plus other semantically related sounds, are then laid out in time for further editing. Here, we have no segmentation but a layering of the selected sounds according to exclusion rules and heuristics.

3.4.5. MATConcat (2004)

The *MATConcat* system [17] (Sturm, 2004a, 2004b), is an open source application in *Matlab* to explore concatenative synthesis. For the moment, units are homogeneous large windows taken out of the database sounds. 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. The application offers many choices of how to handle the case of a non-match (leave a hole, continue the previously selected unit, pick a random unit), and through the use of a large window function on the grains, the result sounds pleasingly smooth, which amounts to a squaring of the circle for concatenative synthesis. *MATConcat* is the first system used to compose two electroacoustic musical works, premiered at ICMC 2004: *Concatenative Variations of a Passage by Mahler*, and *Dedication to George Crumb, American Composer*.

3.5. Group 5: Descriptor analysis with direct selection in real time

This group uses descriptor analysis with a direct local real-time selection of heterogeneous units, without caring for concatenation quality. The local target is given according to a subset of the same descriptors in real time (*MoSievius*, *Musescape* (see section 2.4), *CataRT*, *frelia*).

3.5.1. MoSievius (2003)

The *MoSievius* system [18] (Lazier & Cook, 2003) is an encouraging first attempt to apply unit selection to real-time performance-oriented synthesis with direct intuitive control.

The system is based on sound segments placed in a loop: According to user controlled ranges for some descriptors, a segment is played when its descriptor values lie within the ranges. The descriptor set used contains voicing, energy, spectral flux, spectral centroid, instrument class. This method of content-based retrieval is called *Sound Sieve* and is similar to the *Musescape* system (Tzanetakis, 2003) for music selection (see section 2.4).

3.5.2. CataRT (2005)

The ICMC 2005 workshop on *Audio Mosaicing: Feature-Driven Audio Editing/Synthesis* saw the presentation of the first prototype of a real-time concatenative synthesiser (Schwarz, 2005) called *CataRT*, loosely based on *Caterpillar*. It implements the application of free synthesis as interactive exploration of sound databases (section 1.1) and is in its present state rather close to directed, data-driven granular synthesis (section 3.2.2).

In *CataRT*, the units in the chosen corpus are laid out in a Euclidean descriptor space, made up of pitch, loudness, spectral characteristics, modulation, etc. A (usually 2-dimensional) projection of this space serves as the user interface that displays the units' positions and allows to move a cursor. The units closest to the cursor's position are selected and played at an arbitrary rate. *CataRT* is implemented as a Max/MSP [19] patch using the FTM and *Gabor* extensions [20] (Schnell, Borghesi, Schwarz, Bevilacqua, & Müller, 2005; Schnell & Schwarz, 2005). The sound and descriptor data can be loaded from SDIF files (see section 4.3) containing MPEG-7 descriptors, or can

[15] http://crca.ucsd.edu/~pxiang/research.htm
[16] http://puredata.info
[17] http://www.mat.ucsb.edu/~b.sturm/sand/VLDCMCaR/VLDCMCaR.html
[18] http://soundlab.cs.princeton.edu/research/mosievius
[19] http://www.cycling74.com
[20] http://www.ircam.fr/ftm

be calculated on-the-fly. It is then stored in FTM data structures in memory. An interface to the *Caterpillar* database, to the *freesound* repository (see section 4.3), and other sound databases is planned.

3.5.3. Frelia (2005)

The interactive installation *frelia* [21] by Ali Momeni and Robin Mandel uses sets of uncut sounds from the *freesound* repository (see section 4.3) chosen by the textual description given by the sound creator. The sounds are laid out on two dimensions for the user to choose according to the two principal components of *freesound*'s descriptor space of about 170 dimensions calculated by the *AudioClas* [22] library.

3.6. Group 6: High-level descriptors for targeted or stochastic selection

Here, high-level musical or contextual descriptors are used for targeted or stochastic local selection (*MPEG-7 Audio Mosaics*, *Soundspotter*, *NAG*) without specific handling of concatenation quality.

3.6.1. MPEG-7 Audio Mosaics (2003)

In the introductory tutorial at the DAFx 2003 conference [23] titled *Sound replacement, beat unmixing and audio mosaics: Content-based audio processing with MPEG-7*, Michael Casey and Adam Lindsay showed what they called "creative abuse" of MPEG-7: audio mosaics based on pop songs, 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.

3.6.2. Network Auralization for Gnutella (2003)

Jason Freeman's N.A.G. software (Freeman, 2003) selects snippets of music downloaded from the *Gnutella* p2p network according to the descriptors *search term*, *network bandwidth*, etc. and makes a collage out of them by concatenation.

The descriptors used here are partly content-dependent like the metadata accessed by the search term, and partly context-dependent, i.e. changing from one selection to the next, like the network characteristics.

A similar approach is taken in the forthcoming *iTunes Signature Maker* [24] , which creates a short sonic signature from an iTunes music collection as a collage according to descriptors like play count, rating, last play date, which are again context-dependent descriptors.

3.6.3. SoundSpotter (2004)

Casey's system, implemented in *Pure Data* on a *PostGreSQL* [25] database, performs real-time resynthesis of an audio target from an arbitrary-size database by matching of strings of 8 "sound lexemes", which are basic spectro-temporal constituents of sound. Casey reports that about 60 lexemes are enough to describe, in their various temporal combinations, any sound. By hashing and standard database indexation techniques, highly efficient lookup is possible, even on very large sound databases. Casey (2005) claims that one petabyte or 3000 years of audio can be searched in less than half a second.

3.7. Group 7: Descriptor analysis with fully automatic high-level unit selection

This last group uses descriptor analysis with fully automatic global high-level unit selection and concatenation by path-search unit selection (*Caterpillar*, *Audio Analogies*) or by real-time constraint solving unit selection (*Musical Mosaicing*, *Ringomatic*).

3.7.1. Caterpillar (2000)

Caterpillar, first proposed in (Schwarz, 2000, 2003a, 2003b) and described in detail in (Schwarz, 2004), performs data-driven concatenative musical sound synthesis from large heterogeneous sound databases.

Units are segmented by automatic alignment of music with its score (Orio & Schwarz, 2001) for instrument corpora, and by blind segmentation for free and re-synthesis. In the former case, the solo instrument recordings are split into seminote units, which can then be recombined to *dinotes*, analogous to *diphones* from speech synthesis. The unit boundaries are thus usually within the sustain phase and as such in a stable part of the notes, where concatenation can take place with the least discontinuity. The descriptors are based on the MPEG-7 low-level descriptor set, 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, spectrum, etc.) The database is implemented using the relational SQL database management system *PostGreSQL* for added reliability and flexibility.

The unit selection algorithm is of the path-search type (see section 1.2.7) where a Viterbi algorithm finds the globally optimal 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 by weighted Euclidean distance, including a context around the target, and the *concatenation cost* predicts the quality of the join of two database units by join-point continuity of selected descriptors.

Unit corpora of violin sounds, environmental noises, and speech have been built and used for a variety of sound examples of high-level synthesis and resynthesis of audio.

3.7.2. Talkapillar (2003)

The derived project *Talkapillar* (Kärki, 2003) adapted the *Caterpillar* system for text-to-speech synthesis using specialised phonetic and phonologic descriptors. One of its applications is to recreate the voice of a defunct eminent

[21] http://ali.corpuselectronica.com/projects/frelia/frelia.html
[22] http://audioclas.iua.upf.edu
[23] http://www.elec.qmul.ac.uk/dafx03
[24] http://www.jasonfreeman.net/itsm
[25] http://www.postgresql.org

writer to read one of his texts for which no recordings exist. The goal here is different from fully automatic text-to-speech synthesis: highest speech quality is needed (concerning both sound and expressiveness), manual refinement is allowed.

The role of *Talkapillar* is to give the highest possible automatic support for human decisions and synthesis control, and to select a number of well matching units in a very large base (obtained by automatic alignment) according to high level linguistic descriptors, which reliably predict the low-level acoustic characteristics of the speech units from their grammatical and prosodic context, and emotional and expressive descriptors (Beller, 2004, 2005).

In a further development, this system now allows hybrid concatenation between music and speech by mixing speech and music target specifications and databases, and is applicable to descriptor-driven or context-sensitive voice effects (Beller, Schwarz, Hueber, & Rodet, 2005). [26]

3.7.3. Musical Mosaicing (2001)

Musical Mosaicing, or *Musaicing* (Zils & Pachet, 2001), 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 unit selection as a constraint solving problem (CSP). The set of descriptors used for the selection is: mean pitch (by zero crossing rate), loudness, percussivity, timbre (by spectral distribution). Work on adding more descriptors has picked up again with (Zils & Pachet, 2003, 2004) (see also section 4.2) and is further advanced in section 3.7.4.

3.7.4. Ringomatic (2005)

The work of *Musical Mosaicing* (section 3.7.3) is adapted to real-time interactive high level selection of bars of drum recordings in the recent *Ringomatic* system (Aucouturier & Pachet, 2005). The constraint solving problem (CSP) of Zils and Pachet (2001) is reformulated for the real-time case, where the next bar of drums from a database of recordings of drum playing has to be selected according to local matching constraints and global continuity constraints holding on the previously selected bars.

The local match is defined by four drum-specific descriptors derived by the *EDS* system (see section 4.2): *perceptive energy*, *onset density*, *presence of drums*, *presence of cymbals*. Interaction takes place by analysing a MIDI performance and mapping its energy, density and mean pitch to target drum descriptors. The local constraints derived from these are then balanced with the continuity constraints to choose between reactivity and autonomy of the generated drum accompaniment.

3.7.5. Audio Analogies (2005)

Expressive instrument synthesis from MIDI (trumpet in the examples) is the aim of this project by researchers from the University of Washington and Microsoft Research (Simon, Basu, Salesin, & Agrawala, 2005), achieved by selecting note units by pitch from a sound base constituted by just one solo recording from a practice CD. The result sounds very convincing because of the good quality of the manual segmentation, the globally optimal selection using the Viterbi algorithm as in (Schwarz, 2000), and transformations with a PSOLA algorithm to perfectly attain the target pitch and the duration of each unit.

An interesting point is that the style and the expression of the song chosen as sound base is clearly perceivable in the synthesis result.

[26] Examples can be heard on http://www.ircam.fr/anasyn/concat

4. REMAINING PROBLEMS

This section gives a (necessarily incomplete) selection of the most urgent or interesting problems to work on.

4.1. Segmentation

The segmentation of the source sounds that are to form the database is fundamental because it defines the unit base and thus the whole synthesis output. While phone or note units are clearly defined and automatically segmentable, even more so when the corresponding text or score is available, other source material is less easy to segment. For general sound events, automatic segmentation into *sound objects* in the Schaefferian sense is only at its beginning (Hoskinson & Pai, 2001; Cardle et al., 2003; Jehan, 2004). Also, segmentation of music (used e.g. by Zils and Pachet) is harder to do right because of the complexity of the material. Finally, the best solution would be not to have a fixed segmentation to start from, but to be able to choose the unit's segments on the fly. However, this means that also the unit descriptors' temporal modeling has to be recalculated accordingly (see section 1.2.1), which poses hard problems for efficiency, a possible solution for which is the *scale tree* in (de Cheveigné, 2002).

4.2. Descriptors

Better descriptors are needed for more musical use of concatenative synthesis, and more efficient use for sound synthesis.

Definitely needed is a descriptor for *percussiveness* of a unit. In (Tzanetakis, Essl, & Cook, 2002), this question is answered for musical excerpts, by calibrating automatically extracted descriptors for the *beat strength* to perceptive measurements.

An interesting approach to the definition of new descriptors is the *Extractor Discovery System (EDS)* (Zils & Pachet, 2003, 2004): Here, a genetic algorithm evolves a formula using standard DSP and mathematical building blocks whose fitness is then rated using a cross validation database with data labeled by users. This method was successfully applied to the problem of finding an algorithm to calculate the *perceived intensity* of music.

4.2.1. Musical Descriptors

The recent progress in the establishment of a standard score representation format with *MusicXML* as the most promising candidate, means that we can soon overcome

the limitations of MIDI and make use of the entire information from the score, when available and linked to the units by alignment. This means performing unit selection on a higher level, exploiting musical context information from the score, such as dynamics (crescendo, diminuendo), and better describing the units (e.g. we'd know which units are trills, which ones bear an accent, etc). We can already now derive musical descriptors from an analysis of the score, such as:

Harmony A unit's chord or chord class, and a measure of consonance/dissonance can serve as powerful high-level musical descriptors, that are easy to specify as a target, e.g. in MIDI.

Rhythm Position in the measure, relative weight or accent of the note applies mainly to percussive sounds. This information can partially be derived from the score but should be complemented by beat tracking that analyses the signal for the properties of the percussion sounds.

Musical Structure Future descriptors that express the position or function of a unit within the musical structure of a piece will make accessible for selection the subtle nuances that performers install in the music. This further develops the concept of high-level synthesis (see section 1.1) by giving context information about the musical function of a unit in the piece, such that the selection can choose units that fulfill the same function. For speech synthesis, this technique has had a surprisingly large effect on naturalness (Prudon, 2003).

4.2.2. Evaluation of Descriptor Salience

Advanced standard descriptor sets like MPEG-7 propose tens of descriptors, whose temporal evolution can then be characterised in several parameters. This enormous number of parameters that could be used for selection carries of course incredible redundancies. However, as concatenative synthesis is to be used for musical applications, one can not know in advance, which descriptors will be useful. The aim is to give maximum flexibility to the composer using the system. Most applications only use a very small subset of these descriptors.

For the more precisely defined applications, a systematic evaluation of which descriptors are the most useful for synthesis, would be welcome, similar to the automatic choice of descriptors for instrument classification in (Livshin, Peeters, & Rodet, 2003).

An important open research question is how to map the descriptors we can automatically extract from the sound data to a perceptive similarity space that allows us to obtain distances between units.

4.3. Database and Intellectual Property

The databases used for concatenative synthesis are generally rather small, e.g. 1h 30 in *Caterpillar*. In speech synthesis, 10h are needed for only one mode of speech!

Standard descriptor formats and APIs are not so far away with MPEG-7 and the SDIF Sound Description Interchange Format [27] (Wright, Chaudhary, Freed, Khoury, & Wessel, 1999; Schwarz & Wright, 2000). A common database API would greatly enhance the possibilities of exchange, but it is probably still too early to define it.

Finally, concatenative synthesis from existing song material evokes tough legal questions of intellectual property, sampling and citation practices as evoked by Oswald (1999), Cutler (1994), and Sturm (2006) in this issue, and summarised by John Oswald in (Cutler, 1994) as follows:

If creativity is a field, copyright is the fence.

A welcome initiative is the *freesound* project, [28] a collaboratively built up online database of samples under licensing terms less restrictive than the standard copyright, as provided by the *Creative Commons* [29] family of licenses. Now imagine a transparent net access from a concatenative synthesis system to this sound database, with unit descriptors already calculated [30] — an endless supply of fresh sound material.

4.4. Data-Driven Optimisation of Unit Selection

It should be possible to exploit the data in the database to analyse the natural behaviour of an underlying instrument or sound generation process, which enables us to better predict what is natural in synthesis. The following points are developed in more detail in (Schwarz, 2004).

4.4.1. Learning Distances from the Data

Knowledge about similarity or distance between high-level symbolic descriptors can be obtained from the database by an acoustic distance function, and classification. For speech, with the regular and homogeneous phone units, this is relatively clear (Macon, Cronk, & Wouters, 1998), but for music, the acoustic distance is the first problem: How do we compare different pitches, how units of completely different origins and durations?

4.4.2. Learning Concatenation from the Data

A corpus of recordings of instrumental performances or any other sound generating process can be exploited to learn the concatenation distance function from the data by statistical analysis of pairs of consecutive units in the database. The set of each unit's descriptors defines a point in a high-dimensional descriptor space D. The natural concatenation with the consecutive unit defines a vector to that unit's point in D. The question is now if, given any pair of points in D, we can obtain from this vector field a measure to what degree the two associated units concatenate like if they were consecutive.

The problem of modeling a high-dimensional vector field becomes easier if we restrict the field to clusters of units in a corpus and calculate the distances between all pairs of cluster centres. This will provide us with a concatenation distance matrix between clusters that can be used as a fast lookup table for unit selection. This allows

[27] http://www.ircam.fr/sdif

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

[29] http://creativecommons.org

[30] The license type of each unit should be part of the descriptor set, such that a composer could, e.g. only select units with a license permitting commercial use, if she wants to sell the composition.

us also to use the database for synthesis by modeling the probabilities to go from one cluster of units to the next. This model would prefer, in synthesis, the typical articulations taking place in the database source, or, when left running freely, would generate a sequence of units that recreates the texture of the source sounds.

4.4.3. Learning Weights from the Data

Finally, there is a large corpus of literature about automatically obtaining the weights for the distance functions by search in the weight-space with resynthesis of natural recordings for speech synthesis (Hunt & Black, 1996; Macon et al., 1998). A performance optimised method, applied to singing voice synthesis, is described in (Meron, 1999), and an application in *Talkapillar* is described in (Lannes, 2005).

All these data-driven methods depend on an acoustic or perceptual distance measure that can tell us when two sounds "sound the same". Again, for speech this might be relatively clear, but for music, this is itself a subject of research in musical perception and cognition.

4.5. Real-Time Interactive Selection

Using concatenative synthesis in real-time allows interactive browsing of a sound database. The obvious interaction model of a trajectory through the descriptor space presents the problems of its sparse and uneven population. A more appropriate model might be that of navigation through a graph of clusters of units. However, a good mix of generative and user-driven behaviour of the system has to be found. [31]

Globally optimal unit selection algoritms, that take care of concatenation quality such as Viterbi path search or constraint satisfaction, are inherently non real-time. Real-time synthesis could partially make up for this by allowing transformation of the selected units. This introduces the need for defining a *transformation cost* that predicts the loss of sound quality introduced by this.

Real-time synthesis also places more stress on the efficiency of the selection algorithm, which can be augmented through clustering of the unit database or use of optimised multi-dimensional indices (D'haes, Dyck, & Rodet, 2002, 2003; Roy, Aucouturier, Pachet, & Beurivé, 2005). However, also in the non real-time case, faster algorithms allow for more experimentation and for more parameters to be explored.

4.6. Synthesis

The commonly used simple crossfade concatenation is enough for the first steps of concatenative sound synthesis. Eventually, one would have to apply the findings from speech synthesis about reducing discontinuities (Prudon, 2003) or the recent work by Osaka (2005), or use advanced signal models like additive sinusoidal plus noise, or PSOLA. This leads to *parametric concatenation*, where the units are stored as synthesis parameters that are easier to concatenate before resynthesising.

Going further, hybrid concatenation of units using different signal models promises clear advantages: each type of unit (transient, steady state, noise) could be represented in the most appropriate way for transformations of pitch and duration.

[31] For instance, one particular difficulty is that in real-time synthesis, the duration of a target unit is not known in advance, so that the system must be capable of generating a pleasing stream of database units as long as there is no user input.

5. CONCLUSION

What we tried to show in this article is 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 foster exchange of ideas and experience and help the fledgling community, a mailinglist *concat@ircam.fr* has been created, accessible from (Schwarz, 2006). This site also hosts the online version of this survey of research and musical systems using concatenation which is continually updated.

Professional and multi-media sound synthesis devices or software show a natural drive to make use of the advanced mass storage capacities available today, and of 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.

Where is concatenative sound synthesis now? The musical applications of CSS are just starting to become convincing (Sturm 2004a, see section 3.4.5), and real-time explorative synthesis is around the corner (Schwarz 2005, see section 3.5.2). For high-level synthesis, we stand at the same position speech synthesis stood 10 years ago, with yet too small databases, and many open research questions. The first commercial application (Lindemann 2001, see section 3.2.4) is comparable to the early fixed-inventory diphone speech synthesisers, but its expressivity and real-time capabilities are much more advanced than that.

Data-driven synthesis is now more feasible than ever with the arrival of large sound database schemes. They finally promise to provide large sound corpora in standardised description. It is this constellation that provided the basis for great advancements in speech research: the existence of large speech databases allowed corpus-based linguistics to enhance linguistic knowledge and the performance of speech tools.

Where will concatenative sound synthesis be in a few year's time? To answer this question, we can sneak a look at where speech synthesis is today: Text-to-speech synthesis has, after 15 years of research, now become a technology mature to the extent that all recent commercial speech synthesis systems are concatenative. This success is also due to the database size of up to 10 hours of speech, a size we did not yet reach for musical synthesis.

The hypothesis of high level symbolic synthesis explained in section 1.1 proved true for speech synthesis, when the database is large enough (Prudon, 2003). However, this database size is needed to adequately synthesise

just one "instrument" — the human voice — in just one "neutral" expression. What we set out for with data-driven concatenative sound synthesis is synthesising a multitude of instruments and sound processes, each with its idiosyncratic behaviour. Moreover, research is still at its beginning on multi-emotion or expressive speech synthesis, something we can't do without for music.

6. ACKNOWLEDGEMENTS

Thanks go to Matt Wright, Jean-Philippe Lambert, and Arshia Cont for pointing out interesting sites that (ab)use CSS, to Bob Sturm for the discussions and the beautiful music, to Mikhail Malt for sharing his profound knowledge of the history of electronic music, to all the authors of the research mentioned here for their interesting work in the emerging field of concatenative synthesis, and to Adam Lindsay for bringing people of this field together.

References

Amatriain, X., Bonada, J., Loscos, A., Arcos, J., & Verfaille, V. (2003). Content-based transformations. *Journal of New Music Research*, *32*(1), 95–114.

Aucouturier, J.-J., & Pachet, F. (2005). Ringomatic: A Real-Time Interactive Drummer Using Constraint-Satisfaction and Drum Sound Descriptors. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (pp. 412–419). London, UK.

Aucouturier, J.-J., Pachet, F., & Hanappe, P. (2004). From sound sampling to song sampling. In *Proceedings of the international symposium on music information retrieval (ISMIR).* Barcelona, Spain.

Battier, M. (2001). Laboratori. In J.-J. Nattiez (Ed.), *Enciclopedia della musica* (Vol. I, pp. 404–419). Milan: Einaudi.

Battier, M. (2003). Laboratoires. In J.-J. Nattiez (Ed.), *Musiques. Une encyclopédie pour le XXIe siècle* (Vols. I, Musiques du XXe siècle, pp. 558–574). Paris: Actes Sud, Cité de la musique.

Beller, G. (2004). *Un synthétiseur vocal par sélection d'unités.* Rapport de stage DEA ATIAM, Ircam – Centre Pompidou, Paris, France.

Beller, G. (2005). *La musicalité de la voix parlée.* Maitrise de musique, Université Paris 8, Paris, France.

Beller, G., Schwarz, D., Hueber, T., & Rodet, X. (2005). A hybrid concatenative synthesis system on the intersection of music and speech. In *Journées d'Informatique Musicale (JIM)* (pp. 41–45). MSH Paris Nord, St. Denis, France.

Bonada, J., Celma, O., Loscos, A., Ortola, J., Serra, X., Yoshioka, Y., Kayama, H., Hisaminato, Y., & Kenmochi, H. (2001). Singing voice synthesis combining excitation plus resonance and sinusoidal plus residual models. In *Proceedings of the international computer music conference (icmc).* Havana, Cuba.

Bünger, E. (2003). *Let Them Sing It For You.* Web page. (http://www.sr.se/sing http://www.erikbunger.com/

Cage, J. (1962). *Werkverzeichnis.* New York: Edition Peters.

Cano, P., Fabig, L., Gouyon, F., Koppenberger, M., Loscos, A., & Barbosa, A. (2004). Semi-automatic ambiance generation. In *Proceedings of 7th international conference on digital audio effects.* Naples, Italy.

Cardle, M. (2004). *Automated Sound Editing* (Tech. Rep.). University of Cambridge, UK: Computer Laboratory.

Cardle, M., Brooks, S., & Robinson, P. (2003). Audio and user directed sound synthesis. In *Proceedings of the international computer music conference (icmc).* Singapore.

Casey, M. (2005). Acoustic Lexemes for Real-Time Audio Mosaicing [Workshop]. In A. T. Lindsay (Ed.), *Audio Mosaicing: Feature-Driven Audio Editing/Synthesis.* Barcelona, Spain: International Computer Music Conference (ICMC) workshop. (http://www.icmc2005.org/index.php?selectedPage=120

Chion, M. (1995). *Guide des objets sonores.* Paris, France: Buchet/Chastel.

Codognet, P., & Diaz, D. (2001). Yet another local search method for constraint solving. In *AAAI Symposium.* North Falmouth, Massachusetts.

Cutler, C. (1994). Plunderphonia. *Musicworks*, *60*(Fall), 6–19.

de Cheveigné, A. (2002). Scalable metadata for search, sonification and display. In *International Conference on Auditory Display (ICAD 2002)* (pp. 279–284). Kyoto, Japan.

D'haes, W., Dyck, D. van, & Rodet, X. (2002). An efficient branch and bound seach algorithm for computing k nearest neighbors in a multidimensional vector space. In *Ieee advanced concepts for intelligent vision systems (acivs).* Gent, Belgium.

D'haes, W., Dyck, D. van, & Rodet, X. (2003). PCA-based branch and bound search algorithms for computing K nearest neighbors. *Pattern Recognition Letters*, *24*(9–10), 1437-1451.

DiScipio, A. (2005). Formalization and Intuition in Analogique A et B. In *Proceedings of the international symposium iannis xenakis* (pp. 95–108). Athens, Greece.

Dubnov, S., Bar-Joseph, Z., El-Yaniv, R., Lischinski, D., & Werman, M. (2002). Synthesis of audio sound textures by learning and resampling of wavelet trees. *IEEE Computer Graphics and Applications*, *22*(4), 38–48.

Forney. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, *61*, 268–278.

Freeman, J. (2003). *Network Auralization for Gnutella.* Web page. (http://turbulence.org/Works/freeman http://www.jasonfreeman.net/Catalog/electronic/nag.html

GRAM (Ed.). (1996). *Dictionnaire des arts médiatiques.* Groupe de recherche en arts médiatiques, Université du Québec à Montréal. (http://www.comm.uqam.ca/~GRAM

Hazel, S. (2001). *Soundmosaic.* web page. (http://thalassocracy.org/soundmosaic)

Hoskinson, R., & Pai, D. (2001). Manipulation and

resynthesis with natural grains. In *Proceedings of the International Computer Music Conference (ICMC).* Havana, Cuba.

Hummel, T. A. (2005). Simulation of Human Voice Timbre by Orchestration of Acoustic Music Instruments. In *Proceedings of the International Computer Music Conference (ICMC).* Barcelona, Spain: ICMA.

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 373–376). Atlanta, GA.

Hunter, J. (1999). MPEG7 Behind the Scenes. *D-Lib Magazine*, *5*(9). (http://www.dlib.org/)

Jehan, T. (2004). Event-Synchronous Music Analysis/Synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx).* Naples, Italy.

Kärki, O. (2003). *Système talkapillar.* Unpublished master's thesis, EFREI, Ircam – Centre Pompidou, Paris, France. (Rapport de stage)

Kobayashi, R. (2003). Sound clustering synthesis using spectral data. In *Proceedings of the International Computer Music Conference (ICMC).* Singapore.

Lannes, Y. (2005). *Synthèse de la parole par concaténation d'unités* (Mastère Recherche Signal, Image, Acoustique, Optimisation). Université Toulouse III Paul Sabatier.

Lazier, A., & Cook, P. (2003). MOSIEVIUS: Feature driven interactive audio mosaicing. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)* (pp. 312–317). London, UK.

Lindemann, E. (2001, November). *Musical synthesizer capable of expressive phrasing* [United States Patent]. US Patent 6,316,710.

Lindsay, A. T., Parkes, A. P., & Fitzgerald, R. A. (2003). Description-driven context-sensitive effects. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx).* London, UK.

Livshin, A., Peeters, G., & Rodet, X. (2003). Studies and improvements in automatic classification of musical sound samples. In *Proceedings of the international computer music conference (icmc).* Singapore.

Lomax, K. (1996). The development of a singing synthesiser. In *3èmes journees d'informatique musicale (jim).* Ile de Tatihou, Lower Normandy, France.

Macon, M., Jensen-Link, L., Oliverio, J., Clements, M. A., & George, E. B. (1997a). A singing voice synthesis system based on sinusoidal modeling. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 435–438). Munich, Germany.

Macon, M., Jensen-Link, L., Oliverio, J., Clements, M. A., & George, E. B. (1997b). Concatenation-Based MIDI-to-Singing Voice Synthesis. In *103rd meeting of the audio engineering society.* New York.

Macon, M. W., Cronk, A. E., & Wouters, J. (1998). Generalization and discrimination in tree-structured unit selection. In *Proceedings of the 3rd esca/cocosda international speech synthesis workshop.* Jenolan Caves, Australia.

Manion, M. (1992). *From Tape Loops to Midi: Karlheinz Stockhausen's Forty Years of Electronic Music.* Online article. (http://www.stockhausen.org/tape_loops.html)

Meron, Y. (1999). *High quality singing synthesis using the selection-based synthesis scheme.* Unpublished doctoral dissertation, University of Tokyo.

Orio, N., & Schwarz, D. (2001). Alignment of Monophonic and Polyphonic Music to a Score. In *Proceedings of the International Computer Music Conference (ICMC).* Havana, Cuba.

Osaka, N. (2005). Concatenation and stretch/squeeze of musical instrumental sound using sound morphing. In *Proceedings of the International Computer Music Conference (ICMC).* Barcelona, Spain.

Oswald, J. (1993). *Plexure.* CD. (http://plunderphonics.com/xhtml/xdiscography.html\#plexure

Oswald, J. (1999). *Plunderphonics.* web page. (http://www.plunderphonics.com

Pachet, F., Roy, P., & Cazaly, D. (2000). A combinatorial approach to content-based music selection. *IEEE MultiMedia*, *7*(1), 44–51.

Prudon, R. (2003). *A selection/concatenation TTS synthesis system.* Unpublished doctoral dissertation, LIMSI, Université Paris XI, Orsay, France.

Puckette, M. (2004). Low-Dimensional Parameter Mapping Using Spectral Envelopes. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 406–408). Miami, Florida.

Roads, C. (1988). Introduction to granular synthesis. *Computer Music Journal*, *12*(2), 11–13.

Roads, C. (1996). The computer music tutorial. In (pp. 117–124). Cambridge, Massachusetts: MIT Press.

Roads, C. (2001). *Microsound.* Cambridge, Mass: MIT Press.

Rodet, X. (2002). Synthesis and processing of the singing voice. In *Proceedings of the 1st ieee benelux workshop on model based processing and coding of audio (mpca).* Leuven, Belgium.

Roy, P., Aucouturier, J.-J., Pachet, F., & Beurivé, A. (2005). Exploiting the Tradeoff Between Precision and CPU-time to Speed up Nearest Neighbor Search. In *Proceedings of the international symposium on music information retrieval (ISMIR).* London, UK.

Schaeffer, P. (1966). *Traité des objets musicaux* (1st ed.). Paris, France: Éditions du Seuil.

Schaeffer, P., & Reibel, G. (1967). *Solfège de l'objet sonore.* Paris, France: ORTF. (Reedited as (Schaeffer & Reibel, 1998))

Schaeffer, P., & Reibel, G. (1998). *Solfège de l'objet sonore.* Paris, France: INA Publications–GRM. (Reedition on 3 CDs with booklet of (Schaeffer & Reibel, 1967))

Schnell, N., Borghesi, R., Schwarz, D., Bevilacqua, F., & Müller, R. (2005). FTM—Complex Data Structures for Max. In *Proceedings of the International Computer Music Conference (ICMC).* Barcelona, Spain.

Schnell, N., & Schwarz, D. (2005). Gabor, Multi-Representation Real-Time Analysis/Synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx).* Madrid, Spain.

Schwarz, D. (2000). A System for Data-Driven Concatenative Sound Synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)* (pp. 97–102). Verona, Italy.

Schwarz, D. (2003a). New Developments in Data-Driven Concatenative Sound Synthesis. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 443–446). Singapore.

Schwarz, D. (2003b). The CATERPILLAR System for Data-Driven Concatenative Sound Synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)* (pp. 135–140). London, UK.

Schwarz, D. (2004). *Data-driven concatenative sound synthesis.* Thèse de doctorat, Université Paris 6 – Pierre et Marie Curie, Paris.

Schwarz, D. (2005). Recent Advances in Musical Concatenative Sound Synthesis at Ircam [Workshop]. In A. T. Lindsay (Ed.), *Audio Mosaicing: Feature-Driven Audio Editing/Synthesis.* Barcelona, Spain: International Computer Music Conference (ICMC) workshop. (http://www.icmc2005.org/index.php?selectedPage=120

Schwarz, D. (2006). *Caterpillar.* Web page. (http://recherche.ircam.fr/anasyn/schwarz/thesis

Schwarz, D., & Wright, M. (2000). Extensions and Applications of the SDIF Sound Description Interchange Format. In *Proceedings of the International Computer Music Conference (ICMC)* (pp. 481–484). Berlin, Germany. ()

Simon, I., Basu, S., Salesin, D., & Agrawala, M. (2005). Audio analogies: Creating new music from an existing performance by concatenative synthesis. In *Proceedings of the International Computer Music Conference (ICMC).* Barcelona, Spain.

Sturm, B. L. (2004a). MATConcat: An Application for Exploring Concatenative Sound Synthesis Using MATLAB. In *Proceedings of the International Computer Music Conference (ICMC).* Miami, Florida.

Sturm, B. L. (2004b). MATConcat: An Application for Exploring Concatenative Sound Synthesis Using MATLAB. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx).* Naples, Italy.

Sturm, B. L. (2006). Concatenative sound synthesis and intellectual property: An analysis of the legal issues surrounding the synthesis of novel sounds from copyright-protected work. *Journal of New Music Research*, *35*(1), 23–34. (Special Issue on Audio Mosaicing)

Thom, D., Purnhagen, H., Pfeiffer, S., & MPEG Audio Subgroup, the. (1999, December). *MPEG Audio FAQ.* web page. Maui. (International Organisation for Standardisation, Organisation Internationale de Normalisation, ISO/IEC JTC1/SC29/WG11, N3084, Coding of Moving Pictures and Audio, http://www.tnt.uni-hannover.de/project/mpeg/audio/faq)

Truchet, C., Assayag, G., & Codognet, P. (2001). Visual and adaptive constraint programming in music. In *Proceedings of the International Computer Music Conference (ICMC).* Havana, Cuba.

Tzanetakis, G. (2003). MUSESCAPE: An interactive content-aware music browser. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx).* London, UK.

Tzanetakis, G., Essl, G., & Cook, P. (2002). Human Perception and Computer Extraction of Musical Beat Strength. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)* (pp. 257–261). Hamburg, Germany.

Vinet, H. (2003). The representation levels of music information. In *Computer music modeling and retrieval (CMMR).* Montpellier, France.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, *IT-13*, 260–269.

Wright, M., Chaudhary, A., Freed, A., Khoury, S., & Wessel, D. (1999). Audio Applications of the Sound Description Interchange Format Standard. In *AES 107th convention preprint.* New York, USA.

Xiang, P. (2002). A new scheme for real-time loop music production based on granular similarity and probability control. In *Digital audio effects (dafx)* (pp. 89–92). Hamburg, Germany.

Zils, A., & Pachet, F. (2001). Musical Mosaicing. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx).* Limerick, Ireland.

Zils, A., & Pachet, F. (2003). Extracting automatically the perceived intensity of music titles. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx).* London, UK.

Zils, A., & Pachet, F. (2004). Automatic extraction of music descriptors from acoustic signals using EDS. In *Proceedings of the 116th AES Convention.* Atlanta, GA, USA.