FEATURE EXTRACTION FOR SELF-ADAPTIVE SYNTHESIS

by Risto Holopainen

Aural feedback, in the sense of hearing what one is playing, is essential for any kind of musical performance. If the pitch of a note turns out to sound a bit too high, it is immediately adjusted by the musician. This process can be transferred to the realm of sound synthesis, and to what we will call adaptive synthesis models. The goal, however, will be much more modest than to model the interaction between a real musician and an acoustic instrument. It will also be different, since self-adaptive synthesis is a self-organizing process.

In adaptive synthesis, the use of feature extraction is crucial. The generated sound signal is analysed whilst being produced, and the analysis simultaneously influences the sound generator's control parameters. We will discuss Pierre Schaeffer's thoughts on listening and sound classification, and try to relate them to methods of feature extraction. Finally, we describe one of the simplest possible adaptive synthesis models in detail, and find that it is not so simple, after all.

What is Adaptive Synthesis?

In adaptive synthesis, there is a conceptual tripartition of the synthesis model into a generator, a low-level feature extractor (also called attribute analysis), and a mapping from analysed features back to the generator, although the distinction between these may become blurred. Even a single sine wave oscillator can provide astoundingly rich sonic possibilities, as we will see, since fast modulation can occur, which in effect turns it into an FM instrument.

Synthesis models are often divided into three broad classes: physical, perceptual, and abstract models. As much as in the perceptual models, feature extraction does play an important role in adaptive synthesis. But whereas perceptual models try to develop perceptually relevant and intuitive representations of synthesis parameters, adaptive synthesis escapes such goals. It differs also from physical modeling, as it does not try to model any known physical system.

Adaptive synthesis may be viewed as taking an intermediate position between an instrument and a tool for algorithmic composition. It is common to refer to such digital instruments as 'composed instruments' (Schnell & Battier 2002), since the instrument incorporates a bit of a score function, and is often constructed with a particular and limited musical purpose in mind. In the extreme, the instrument becomes synonymous with the composition. This clearly applies to adaptive synthesis models, in the sense that for a sufficiently elaborated model, given fixed parameter values, it might produce a musical sequence that is varied over a lenghty time span.

Expectations of what an instrument should be vary strongly amongst practitioners of digital instruments. Magnusson and Hurtado (2007) pose the relevant question of how the indeterminacy or "entropy" of a digital instrument is experienced by the user. They found that for most users, some indeterminacy can be tolerated and even welcome in an acoustic instrument, but in digital instruments, it is often seen as a flaw. The exception came from some practitioners who deliberately sought out faults or software glitches (cf. the term 'glitch' as a musical genre label). Adaptive synthesis can be particularly prone to sluggish reactions or very long transients to parameter changes, with certain consequences for live performance.

Aural feedback is indispensable in musical performance. Additional haptic[1] feedback comes for free in acoustic instruments, and is sometimes incorporated in controllers for digital instruments as well. Musicians continuously adjust aspects of their playing such as embouchure in a wind instrument, or speed, pressure and angle of bowing in a string instrument. This adjustment happens mostly unconsciously in an expert performer, as a response to the aural and haptic feedback from the instrument. Obviously, this is in close analogy with an adaptive synthesis system.

Various forms of interactive computer music have been developed, some of which clearly utilize a strategy of adaptive synthesis. George Lewis describes his interactive work with computers in terms of a negotiation, a context highly suitable for improvisation. The use of feature extraction is crucial. Such features as volume, velocity, durations of sound and inter-onset duration, pitch and several others are extracted and used to control the programs' response (Lewis 1999).

It is interesting to note that a musical automaton need not be a completely autonomous device. In order to get any ongoing musically interesting response, the easiest way would be to regularly supply new input to the machine. Sinan Bökesoy developed adaptive synthesis with an emphasis on the emergence of multiple temporal levels of change, using granular synthesis (Bökesoy 2007). According to Bökesoy's description, these systems are viewed in terms of dissipative structures, that need to be kept in motion by an incoming energy supply; the 'energy' for instance being user input.

Agostino Di Scipio has done extensive work on a variant of adaptive synthesis that involves the performance space as a part of the system, and which has resulted in a series of works entitled *audible ecosystems* (Di Scipio 2003). Ambient sound is picked up by microphones, analyzed and pro-

1 Examples of haptic feedback in digital instruments could be a knob or slider that yields resistance to rapid motion. The knob is not merely a sensing device, but one that produces a mechanical force.

cessed in a computer, and sent out via loudspeakers in the same room. The processing often has the purpose of counterbalancing features in the sound. If the amplitude is low, gain is increased, if too high, it is reduced. These systems often involve several interconnected feedback loops, which only need some ambient noise to get started, and that are able to sustain a sonic process on their own. It should be noted, that there is no general agreement on terminology, so what is called adaptive synthesis in this paper, may in other places go under such names as autopoiesis[2] or self-organizing systems.

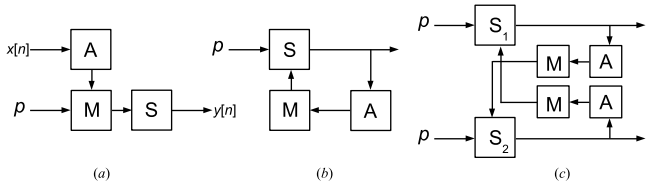

Fig. 1. S: synthesis module, A: attribute analysis, M: mapping. The synthesis model is also controlled by user supplied parameters, p. Left: a recursive structure for selfadaptive synthesis. Right: crossadaptive synthesis.

The way the components of an adaptive synthesis model are connected is of great importance. A general scheme of analysis—modification—resynthesis (fig. 1*a*) has been frequently employed in additive synthesis and linear predictive coding. Remove the input, and connect the output to the attribute analyzer, and we have the self-adaptive model (fig. 1*b*). More complex structures are possible, such as cross-adaptive synthesis (fig. 1*c*). In a nested model, the generator itself is an adaptive synthesis model, but such models are likely to become very complex.

Adaptive Synthesis Models are Dynamical Systems

What does a typical output from an autonomous adaptive synthesis system sound like? Quite naturally, it depends on the details of the particular synthesis model in question, but, deeming from a few preliminary experiments, some general traits can be distinguished. Typical findings include:

- Self-adaptive synthesis may exhibit a wide range of behaviour, including startup transients gravitating towards a relaxed state; subtle and irregular variations; in some cases sudden and unexpected changes after periods of stability. Under certain circumstances, these systems may become chaotic.

- Prominent hysteresis effects are likely to occur; i.e. there may be a general sluggishness in the response to actions, and the response to the same control parameter settings may differ depending on the immediate past history of the system. The window length of the signal attribute analysis is likely to have a strong impact on the types of behaviour that the model will display.

- There is reason to believe that these systems easily become computationally irreducible, that is to say, one cannot predict the output at a given time in the future from the initial conditions and parameter values, but one has to run the entire calculation up to that moment. These models also provide striking examples of emergent phenomena. Decisions on sample level have implications on much higher temporal levels, and there is very little to be inferred about the system's actual behaviour only from its rules.

Linear source-resonance models may capture the dynamics of many acoustic musical instruments rather well, but on closer inspection nonlinearities are often found in the system. Substituting nonlinearities for linearities in a dynamic system may imply drastic qualitative changes, depending on the particular nonlinearity and where it is inserted. Memory-less nonlinear systems also do not provide the same richness of behaviour that awaits us in a dynamic nonlinear system.

Distorted amplifiers and distortion effects are often modeled as memoryless devices. When they are modeled as dynamical systems, it is typically done by Volterra series expansions, i.e. polynomials of the signal variable including delays (Schattschneider and Zölzer 1999). On the other hand, a sufficiently general time-invariant nonlinear filter may be described by the equation

$$\textbf{(1)}\quad y_n = f(x_n,\ x_{n-1},\ \ldots,\ x_{n-K},\ y_{n-1},\ \ldots,\ y_{n-L})$$

where $f(\cdot)$ is any suitable nonlinear function. Recursive nonlinear filters are not so easily represented by Volterra series, whereas their implementation follows directly from the equation (1). In order to describe the behaviour of some of these nonlinear filters, one may resort to study their responses to impulses of varying amplitude, in effect treating them as maps (Holopainen 2007). In fact, virtually all of the signal attribute extractors that are so important in adaptive synthesis, may be thought of as (non-recursive) nonlinear filters.

Maps (iterated functions) may serve as a framework for the study of adaptive synthesis. The essence of adaptive synthesis is an iterated loop of synthesis _ attribute extraction _ synthesis parameter adjustment _ synthesis. This loop is generally traversed once per sample, although the rates of attribute extraction and parameter updates may be slower. Maps with delayed variables can be dealt with by transforming them into higher dimensional delay-less systems.

If the adaptive synthesis models are to be studied as iterated maps, then it is obvious that there is a large delay

2 Literally 'self creation', as in a system capable of maintaining itself while exchanging matter or energy with its environment, such as a living cell.

caused by the attribute analysis. Even for short analysis windows, hundreds of samples of delay could result, or in more realistic cases, tens of thousands. The reason for this should be seen in relation to auditory perception, where incoming information needs to be integrated over a temporal window of some duration.

Auditory Perception and Feature Extraction

Applications of feature extraction arise in several fields, such as clinical diagnosis of voice disorders (Hadjitodorov and Mitev 2002), music information retrieval, score following and automated transcription, in adaptive audio effects and in sound synthesis. In concatenative synthesis[3], a target sound is recreated by matching perceptual features to sound fragments from a database of other sounds. While a large and varied database and efficient matching criteria allows verisimilar reconstruction of the target sound, a more restricted database can be used in a way reminiscent of cross synthesis—say, to articulate the morphological shape of Rite of Spring with a collection of saxophone samples (Sturm 2004). Feature-based synthesis (Hoffman & Cook 2006) also tries to match a feature vector against sounds produced by a synthesizer in a way that is comparable to that of concatenative synthesis.

Adaptive audio effects have the feature extraction in common with adaptive synthesis. Dynamic processing (companders, noise gates, etc) are the earliest known examples of adaptive effects. The sound of an auto-tuner applied to vocals is often used and abused, and should also be familiar. Another example is selective time stretching, where for instance transients can be left in the original time scale, whereas quasi-stationary portions of the sound are stretched (Verfaille and Arfib 2001). Adaptive effects can provide a great coherence between the input sound and the applied effect.

As mentioned above, the length of the temporal frame that is analysed is an influential factor in determining how a self-adaptive synthesis instrument turns out to function. Pitch perception is not possible for sound bursts lasting shorter that at least one or a few periods of the waveform. As the length is increased, the pitch percept gradually becomes more accurate, to a certain limit. Likewise, loudness perception does not follow the immediate pressure variations of the waveform, but takes an average over an extended time window. The same obviously holds for any perceptual attribute.

Studies of detection thresholds for short sound bursts, sinusoidal or noise, have shown that there is a time-intensity trade; the shorter the sound, the louder it needs to be in order to get detected. If both time and intensity of the detection threshold are plotted on logarithmic scales, with time on the abscissa, then the slope of the curve is about -3/4 in a range of durations from 5 periods to 500 ms (Eddins and Green 1995). It has been suggested that a mechanism known as temporal integration could predict these results. In essence, temporal integration is modeled as the convolution of the input signal with a filter impulse response, which is often assumed to be exponentially decaying. This can be directly compared to a common implementation of RMS amplitude extraction, which uses a one pole filter that has the exponentially decaying impulse response. There are, however, other possible implementations of RMS amplitude extraction, such as calculation from a moving average filter, where the impulse response has the shape of a rectangular box. In adaptive synthesis, the goal is not an accurate modeling of the auditory system, yet it may be convenient to have signal attribute extractors that can provide perceptually salient information.

Higher level perceptual attributes, such as vibrato and tremolo, are in fact second order attributes, or descriptors of the time varying curves of the primary attributes. Hence these higher level attributes would always pertain to a longer time scale, and their perception would presumably involve more neural processing, and perhaps the combination of several lower level attributes. Higher order attributes can be extracted from these basic descriptors, by analyzing them further.

Listening with Schaeffer's Typomorphology

Knowledge of the correlations between physical signal and auditory percept will help informing the choice of signal attributes to analyze. In this context, the pioneering work of Schaeffer (1966) may suggest new criteria to analyze (for an excellent short introduction to Schaeffer's typomorphology, see Thoresen 2007). This poses some difficulties however, since Schaeffer's typomorphology is a phenomenologically oriented framework for classifying all conceivable types of sound fragments experienced as a unit. But an exact interpretation of various Schaefferian terms is yet to be made, not to mention a practical signal processing implementation. On the other hand, Schaeffer is quite clear about the openness of the classification system. There is no attempt to arrive at a final classification of a sound object in a specific category. "Nous pensons en effet que le principe de notre classification permet d'assigner au même objet diverses cases selon l'intention d'écoute. La recherche d'une typologie " absolue " est illusoire" [In fact, we think that the

3 Concatenative synthesis was first used in speech synthesis. In the last decade, its utility in musical applications has been seen, where its capability of realistic rendering of any sound rivals that of physical modeling.

principle of our classification permits us to assign different categories to the same sound object, depending on the listening intention. The search for an 'absolute' typology is illusory.] (TOM: 433)[4].

With the metaphor of a loft where heaps of unsorted objects have accumulated, Schaeffer brings up the question of how to choose relevant sorting criteria. Should the violin be placed together with the logs just because both are made of wood, or should small objects be grouped together in one place and large ones in another; or should one despair and give up, the loft being a place where one hides unsortable things away (TOM: 429)? Similarly, before deciding upon a typology of sounds, the sorting criteria have to be chosen. Because this sorting should be carried out according to aural qualities only, one has to focus on the sound as such, and try to ignore what one knows of its causality and meaning.

Reduced listening, *écoute reduite,* is a central concept in Schaeffer's system of thought, but it is a concept that some have found difficult to assimilate. In reduced listening, the attention is focused on the perceived sound as such, with any knowledge of its context, its meaning and its causes bracketed out. A total cutting-out of any anecdotal associations in the sound may be an unattainable ideal, but that is not to say that reduced listening is impossible. On the contrary, as Schaeffer discovered with the aid of a record player with a closed groove, the *sillon fermé,* repeated listening to a short sound fragment soon exhausted any possible interest in its meaning and causes. Reduced listening is a matter of turning one's attention towards morphological traits of the sound, a habit that is actually quickly acquired in the electroacoustic studio.

Schaeffer's typology and morphology is based on the concept of the sonorous object, *l'objet sonore,* which is constituted in the act of reduced listening. A first stage would be to identify the object, which may be delimited by discontinuities in the sound. Next, a suitable category in the typological scheme would be identified, and finally finer details could be described through concepts from the morphology. Classifying and describing sound is challenging for many reasons, one of them being that depending on the type of sound, various traits do or do not exist. A vibrato requires pitch, for instance, and the degree of irregularity of this vibrato certainly requires the vibrato to be present in the first place. It should be noted that the same challenge may have to be met in automated signal attribute extraction, as one may imagine certain attributes that are not always applicable as descriptors of the sound.

One inventive aspect of Schaeffer's typology, is to sort sounds according to their length. Short, percussive or impulsive sounds are classified in one category, as are sounds of medium duration, such as typical musical notes; and sounds of prolonged duration form a third group that is further divided into constant prolonged sounds, variable sounds, and iterated sounds. Other criteria sort sounds according to their energetic articulation *(entretien),* and sonic substance, related to the spectral distribution *(masse).* As in the metaphor of the attic, the first sorting takes a fresh look at things and divides them into broad and somewhat surprising categories. But a prerequisite is that one applies reduced listening when characterizing the sound objects—otherwise one risks relying on prejudice, in effect categorizing objects by other criteria than strictly aural ones.

As Godøy (2006) has pointed out, the categories may not have such a lofty origin after all, but may have a grounding in common gestural types. According to this line of thought, Schaeffer's concepts can be understood in terms of embodied cognition, where virtually all domains of thinking and perception are thought to be related to images of movement. Embodied cognition proposes that perception is not merely carried out as a processing of sensory data, but rather as a re-enactment of whatever is perceived. Some of the evidence for this view comes from neurological studies that have found activation in motor areas of the brain when subjects are asked to imagine music. At least, it is obvious that impulsive, sustained and iterative types can easily be put into correspondence with gestures, but according to Godøy, most or all of the typology and morphology may be matched to gestures—the point is that "there is a gesture component embedded in Schaeffer's conceptual apparatus which is on a more general and basic level than that of everyday causal listening, i.e. not on a level that the principle of reduced listening is supposed to lead us away from" (Godøy 2006: 154).

Relating Schaeffer's Criteria to Signal Analysis

If Schaeffer's typomorphology is to serve as inspiration for new signal attributes to analyze in the context of adaptive synthesis, one is practically restricted to criteria that are independent of the length of the sound object. This leaves criteria that deals with mass, granularity and gait *(allure).* The reason for this is that the signal analysis doesn't operate on segmented sound fragments, but on a continuous stream. But even the perception of a static object is in flux. Godøy (2006) brings up the example of the Necker cube, with the spontaneous changes of its perceived spatial arrangement. An example more related to music would be the fact that a pulse stream of identical impulses is spontaneously grouped into twos or threes, no other elementary grouping is possible. The perception of a sound that is ex-

4 My translations of the quotes from Schaeffer should not necessarily be taken as suggestions of how certain much discussed terms should be rendered in English. Henceforth, references to Schaeffer's Traité des Objets Musicaux will be abbreviated TOM.

perienced as completely static (say, a sinusoid), would presumably also be faced with the effects of fatigue, although it is unlikely that changes in attention level would contribute to a segmentation of the sound.

In contrast, a changing sound stream has points where it is more natural to cut, so a segmentation into sound objects results. Such a segmented sound stream would, at least in theory, be susceptible to analysis in Schaefferian terms, even through machine listening.

Some of the criteria that Schaeffer introduced seem to be easier than others to translate into signal processing routines. Consider the criterion of sonic substance, which is the basis for the seven-step scale from pure tone to noise (TOM: 518). Attribute extractors that recognize sinusoids and white noise as opposite extremes of some continuum do exist. But here, as in an attempt to analyze a sound by ear, it is the intermediate positions that leave most room for interpretation. For instance, a measure of sensory dissonance (Sethares 2005) may have white noise correspond to maximal dissonance, and for a signal of corresponding RMS amplitude, a sinusoid would have minimal dissonance (silence has maximal consonance according to this model). Another measure that would give the desired answers is the spectral crest factor. A single spike, which occurs in the spectrum of a sine tone, has high crest factor, while a flat spectrum has low crest factor. But the spectral crest factor is not a reliable measure of a sound's placement in the sine-to-noise continuum, because an impulse (in the time domain) and a chirp—a sinusoid swept from Nyquist to DC—both share the lowest spectral crest factors along with white noise. It would be possible to define other attribute extractors, which would differentiate between sinusoids and white noise, and which would meet different qualifications in the middle region. This parallels the fact, that one can easily construct several different synthesis models that can generate a range of sounds from a sine tone to noise at the control of a single parameter.

Some of Schaeffer's criteria operate on a very abstract level, and concern such aspects as the mode and degree of variation in the sound. When arriving at a definition of the term *criteria,* Schaeffer emphasizes that these are properties of the perceived sonorous object, and not measurable properties of the physical object. In the days when the *traité* was written, acousticians were well aware of the differences between physical frequency and perceived pitch, between amplitude and intensity, and between duration and perceived temporal length. With the current proliferation of signal attribute analysis methods, one should not forget this important distinction between physical attribute and perceived quality. This relationship, which goes under the name *anamorphose* in Schaeffer's terminology, is characterized by nonlinear correlations, a kind of warping.

Not only the anamorphoses pose difficulties when one is looking for accurately defined correspondences between stimuli and perception; it gets even worse for the fact that several different acoustic dimensions may sometimes contribute to the same morphological traits. "Remarquons aussi que certains critères que nous avons définis ne correspondent à aucun paramètre acoustique simple : c'est le cas du grain, de l'épaisseur, du volume, de l'allure, qu'il est pourtant facile d'isoler, de désigner à l'attention d'un auditeur." [Let us also remark that certain criteria that we have defined do not correspond to any simple acoustic parameter. This holds for the grain, the thickness/depth, the volume, the gait, which it is nevertheless easy to isolate, and to direct a listener's attention to.] (TOM: 502). On the signal processing side, a similar abstraction of attribute analysis is sometimes carried out, as when an attribute is submitted to further statistical analysis. It is common to analyze the mean and variance of an attribute, and many other measures of statistical distribution may be applied as well. However, it is possible that several quite different sounds share a common gait, which is realized as a cyclic, more or less regular variation in some unspecified acoustic dimension. In such a diffuse case, one could arguably define a signal attribute corresponding to the concept of gait as a rate, depth and regularity of variation of the set of all of the other more elementary level signal attributes. But it is crucial to note, that as one decides upon one particular definition of gait, along with the exact details of its signal processing implementation, one has in effect chosen a particular 'listening strategy' (if the term applies to machine listening) among several conceivable ones.

It is not evident that the démarche of turning the typomorphology into rigorously defined signal attributes is the most adequate one. In fact, the opposite strategy, which starts from arbitrary signal attributes and tries to find what, if any, perceptual attribute matches it, may turn out to be a more fruitful approach. Musical analysis in general operates by reducing the overwhelmingly rich information content of the music, be it in notated or recorded form, to a few salient facts. In a similar way, signal attribute analysis works by reducing information to a measure that will hopefully reveal something interesting about the sound.

A digital signal may be submitted to many kinds of processing, which in some sense will reduce its information content (in the information theory sense, this is already implied by smoothing out a signal by averaging over a window). Some of these processed signals will stand in a more straightforward relationship to perceptual dimensions of sound than others. Arguably, it is irrelevant to know the name or precise nature of a perceptual attribute that is closely correlated to a signal attribute. What counts is, whether one is able to reliably distinguish sounds that have a high value of this signal attribute from those having a low value.

To give a concrete example, consider spectral irregularity, which is a measure of the jaggedness of the spectral envelope. It is computed by comparing the spectrum to a three-point smoothed version of itself, where 0 corresponds to a smooth spectrum and 1 corresponds to a maximally irregular spectrum (Beauchamp 2007: 55-58). Beauchamp notes that "... spectral irregularity appears to have a profound effect on a sound's timbre" (op.cit: 58). In listening tests, many recorded instrument sounds have readily been distinguished from the same sounds re-synthesized with smoothed spectral envelopes. Beauchamp concludes that: "Still, despite its obvious importance, no particular perceptual attribute has been found to correspond with spectral irregularity" (ibid).

Case Study: the Sinusoid Generator

One of the simplest possible synthesis models would be a generator that produces a single sine tone, with a parameter controlling its frequency. When the waveform is a sinusoid, it is straightforward to analyze its frequency, as it corresponds to half the number of zero crossings per second. Trevor Wishart (1994) developed several original audio effects that depend on detection of the signal's zero crossings. A segment between two zero crossings going in the same direction is called a waveset, which is the basic unit for various signal operations. Indeed, these effects can be regarded as a kind of adaptive audio effects, since they utilize a rudimentary form of signal analysis.

If the frequency of a sinusoid is known to be constant, it can be calculated from a minimum of two consecutive zero crossings. Assuming that the waveform may have a variable number of zero crossings per period, then zero crossing rate (ZCR) is no longer a reliable momentaneous frequency estimator. But in this case, we know that the wave shape will always be a sine wave, so the ZCR is a very suitable attribute to analyze.

A simple algorithm for ZCR calculation consists of two zero crossing detectors, one operating at the current sample, and the other delayed by D samples. If there is a zero crossing between x_n and x_{n-1}, an accumulator variable c is incremented by one, and if there is a zero crossing between x_{n-D} and x_{n-D-1}, c is decremented by one. The averaged zero crossing rate is c/D.

Now, the generator producing a sinusoid at momentaneous frequency f_n,

(2)
$$x_n = \sin(\theta_n),$$
$$\theta_n = \theta_{n-1} + 2\pi f_n / f_s$$

operating at sample rate f_s is connected to the ZCR analyzer. The output is

(3)
$$z_n = \frac{1}{2} ZCR(x_n)$$

which is divided by 2 since the zero crossing rate indicates crossings in both directions, hence twice the number of cycles per second. Next we will need a mapping M from the analyzed frequency z_n to the frequency f_n in the generator.

The user specifies an initial frequency F of the oscillator, which is assumed to be constant over the sound's duration. There are also user controls for a coupling strength C, to be applied in the mapping function, and analysis window length in seconds, which are also constant. Then the mapping becomes a function of three variables:

(4)
$$f_n = M(z_n,\ C,\ F)$$

In this simple example, all of the surprise lies in the choice of mapping. For several simple mappings, it turns out that the oscillator typically starts out at the initial frequency F, and gradually makes a transition to a new frequency that remains stable for the rest of the sound's duration. Maps having this behaviour include the linear map,

(5)
$$f_n = (1 + Cz_n)F$$

and the quadratic map,

(6)
$$f_n = (1 + Cz_n^2)F$$

both of which produce a rising glissando for $C > 0$, and a falling glissando with an additional decaying vibrato for $C < 0$. For large absolute values of C, aliasing effects may occur, but generally, either the system gradually stabilizes at one frequency, or it approaches a limit cycle resulting in a kind of tremolo.

Slightly more complex results can be obtained with the sine map:

(7)
$$f_n = (1 + \sin(Cz_n))F$$

Typical for this map is a decaying vibrato, at a frequency that is inverse to the analysis window length. Since the ongoing frequency adjustment is simply frequency modulation, it is not too surprising to find a few of the lowest harmonic partials present in the sound (cf fig. 2) [SOUND EXAMPLE 1].[5]

5 Ed. Note: Sound Examples may be found on the Sonic Ideas website, http://www.cmmas.org/sonicideas/

Figure 2. Sine map. F = 880 hz, C = 4, window length 0.2 seconds. It takes about one minute for the vibrato to decay to an imperceptible level. This constant Q filterbank analysis with semitone spacing of analysis bands shows the first 20 seconds.

Clearly the maps (5-7) introduce a bias to the initial frequency, so that the final frequency that is obtained if the system eventually stabilizes is not the same as the initial frequency *F*. A simple way to remove a bias is to high-pass filter the signal. A DC-blocking filter can be incorporated in the mapping as follows:

(8)

$$y = DC(z_n)$$
$$f_c = (1 + Cy)F$$

Here, DC is a second order recursive filter, that not only removes the DC component, but also attenuates frequencies near $f_s/2$. The DC-blocker has the transfer function:

(9)

$$H(z) = \frac{1 - z^{-2}}{1 - Bz^{-2}}$$

where *B* is normally set to 0.995. As long as the coupling is small, this mapping is indeed well behaved. The oscillator starts at a frequency that is higher or lower than *F*, depending on the sign of C, and approaches F in a stepwise fashion. *F*or very large values, more bizarre things happen. Here, the analysis length is of vital importance, and different lengths will cause great qualitative differences in the sound. If the window is sufficiently small, the system will settle into a cycle of repeated distinct pitches. With slightly longer windows, irregular melodic patterns may result, that balance repetition and novelty in a striking way. The beginning of one such tune is shown in fig. 3 [SOUND EXAMPLE 2]. It does not stabilize into a simple pattern, at least not in the first few minutes.

This algorithm, the sine generator with zero crossing rate analysis and a mapping that is essentially just a highpass filter, provides clear evidence of the fact that even the most simple adaptive synthesis model yields unpredictable behaviour in a timespan many orders of magnitude longer than that of the analysis window.

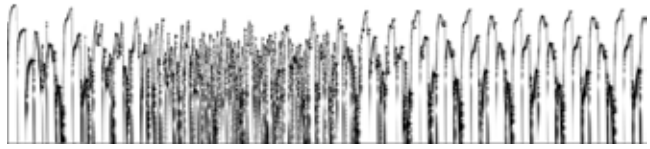

Fig. 3. Mapping with DC-blocker. F = 440hz. C = 8F, window length 0.3 seconds. Almost one half minute is analyzed with the constant-Q filter bank.

In this sinusoidal model, the zero crossing rate corresponds to average frequency. But other frequency or pitch analyzers could be used. Momentaneous frequency can be recovered from a slowly modulated sinusoid by calculating the derivative of the momentaneous phase angle, which can be found with the help of the Hilbert transform. A convenient formula is:

(10)

$$f(t) = \frac{1}{2\pi A^2(t)}(x(t)y'(t) - y(t)x'(t))$$

where $x(t)$ and $y(t)$ is a Hilbert transform pair, and $A(t)$ is the momentaneous amplitude.

The momentaneous frequency can be substituted for the ZCR in the self-adaptive system above. To get comparable results, the momentaneous frequency needs to be smoothed over a temporal window of a length corresponding to that of the ZCR. If the oscillator produces a complex tone instead of a sinusoid, it is preferable to analyze pitch instead of frequency.

Conclusions

Self-adaptive synthesis may be seen as a metaphor of a musician's interaction with an instrument. Depending on traits of the generated sound, the synthesis parameters are continuously adjusted. But the choice of an appropriate mapping is non-trivial, and the output may be very different from anything a musician would produce.

This study of adaptive synthesis does not involve real-time interaction, which may seem a strange priority. After all, musicians possess a unique knowledge and can bring life even to the dullest sounding synthesis technique. Besides, in an interactive situation, the musician can decide when the adaptive synthesis instrument needs new input, if it happens to get stuck in a less interesting corner of its sonic potentialities. But adaptive systems easily become intractably complicated. If the complexity reaches a certain level, very little can be predicted about the system's future behaviour. Thus it should make sense, as a first step, to study these instruments alone, undisturbed by a musician's input, in order to gain at least a basic understanding of their dynamics.

If the exact sonic result is the least important, adaptive synthesis may be used to generate raw material that may or may not be selected for inclusion in a composition. Alternatively, for more adventurous artists, the unpredictability may be essential. Serendipitous discoveries await the explorer all around.

While the three main components of adaptive synthesis models—the generator, the feature extractor, and the mapping—are probably of equal importance, we have focused on strategies for feature extraction and choices of mapping. Schaeffer's typomorphology may not be suited for a direct translation into signal processing routines if one wishes to respect the underlying methodology, which proceeds by reduced listening. On the other hand, in a footnote Thoresen suggested precisely that: "The analysis of sound based on reductive listening is that aspect of musical analysis that best would render itself for an automatic analysis [...]" (Thoresen 2007: 132). This is true, if only because other aspects, such as source recognition and semantic analysis are harder problems. However, Thoresen concludes that the point is "the training of aural consciousness itself" (ibid), which indeed remains an indispensable goal. We have also noted that problems of taxonomy corresponding to those that Schaeffer faced do arise in any attempt to model more complicated signal attributes.

To play an instrument gives rise to a very different experience from merely listening to it. The same goes for programming 'composed instruments'. In this case, one is well aware of the sound's causality—if not, one tries to find out by experiments—and this is, of course, the opposite of reduced listening. "La curiosité scientifique, bien que mettant en jeu des connaissances hautement élaborées, poursuit un but fondamentalement semblable à celui de la perception spontanée de l'événement" [Scientific curiosity, while involving highly elaborated knowledge, pursues a goal fundamentally similar to that of the spontaneous perception of an event.] (TOM: 115).

References

Beauchamp, J. (2007). Analysis and Synthesis of Musical Instrument Sounds. In Beauchamp, J. (ed): Analysis, Synthesis, and Perception of Musical Sounds. Springer.

Bökesoy, S. (2007). Synthesis of a Macro Sound Structure within a Self-Organizing System. Proc. of the Int. Conf. on Digital Audio Effects (DAfX-07). Bordeaux, France.

Di Scipio, A. (2003). 'Sound is the interface': from interactive to ecosystemic signal processing. Organised Sound 8(3): 269-277.

Eddins, D. and Green, D. (1995). Temporal Integration and Temporal Resolution. In Moore, B. C. J. (ed): Hearing. Handbook of Perception and Cognition. Second edition. Academic Press.

Godøy, R. I. (2006). Gestural-Sonorous Objects: embodied extensions of Schaeffer's conceptual apparatus. Organised Sound 11 (2), pp. 149-157.

Hadjitodorov, S. and Mitev, P. (2002). A Computer System for Acoustic Analysis of Pathological Voices and Laryngeal Diseases Screening. Medical Engineering & Physics 24 (2002), pp. 419-429.

Hoffman, M. and Cook, P. (2006). Feature-Based Synthesis for Sonification and Psychoacoustic Research. Proceedings of the 12th International Conference on Auditory Display, London, UK, June 20-23, 2006.

Holopainen, R. (2007). Nonlinear Filters. Proceedings of the ICMC 2007, Vol 1, pp. 283-286. Copenhagen, Denmark.

Lewis, G. (1999). Interacting with Latter-Day Musical Automata. Contemporary Music Review 1999, Vol. 18, Part 3, pp. 99-112.

Magnusson, T. and Hurtado, E. (2007). The Acoustic, the Digital, and the Body: A Survey on Musical Instruments. Proceedings of the 2007 Conference on New Interfaces for Musical Expression (NIME07), New York, USA.

Schaeffer, P. (1966). Traité des Objets Musicaux. Paris: Edition du Seuil.

Sethares, W. (2005). Tuning, Timbre, Spectrum, Scale. Second edition. Springer.

Schattschneider, J. and Zölzer, U. (1999). Discrete-time Models for Nonlinear Audio Systems. Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99), Trondheim, December 9-11, 1999.

Schnell, N. and Battier, M. (2002). Introducing Composed Instruments, Technical and Musicological Implications. Proceedings of the 2002 Conference on New Instruments for Musical Expression (NIME-02), Dublin, Ireland, May 24-26, 2002

Sturm, B. (2004). Matconcat: an Application for Exploring Concatenative Sound Synthesis using Matlab. Proc. of the 7th Int. Conference on Digital Audio Effects (DAFx'04). Naples, Italy, Oct. 5-8, 2004. pp. 323-326.

Thoresen, L. (2007). Spectromorphological Analysis of Sound Objects: an Adaptation of Pierre Schaeffer's Typomorphology. Organised Sound 12 (2). pp. 129-141.

Verfaille, V., and Arfib, D. (2001). A-DAFX: Adaptive Digital Audio Effects. Proc. of the COST-G6 Conf. on Digital Audio Effects (DAFX-01). Limerick, Ireland.

Wishart, T. (1994). Audible Design. A Plain and Easy Introduction to Practical Sound Composition. Orpheus the Pantomime.