Time-Shifting and Transposition of Sampled Sound with a Real-Time Granulation Technique

Barry Truax
Dept. of Communication & School for the Contemporary Arts
Simon Fraser University, Burnaby, B.C., Canada V5A 1S6

ABSTRACT

The author describes a digital signal processing technique using short grains of sampled sound, generally less than 50 ms, that extends the sound's duration without altering its pitch. In addition, individual voices of the stochastically sampled sound may be transposed to a variety of harmonic frequencies. The psychoacoustic implications of the technique, such as the magnification of instantaneous resonances and the perception of increased volume, are discussed, as well as compositional experience that links the inner complexity of the sound to the complexity of the external world.

1 Introduction

Since 1986 I have been working with the technique of granular synthesis [Roads, 1978, 1988, 1991] and since 1987 with the granulation of sampled sound in real-time [Truax, 1988] using the microprogrammable DMX-1000 Digital Signal Processor [Wallruff, 1979]. Briefly, this technique produces complex sounds by the generation of high densities (e.g. 100 - 2000 events/sec) of small "grains" on the order of magnitude of 10 - 50 ms duration. The content of the grain itself can be a fixed waveform, simple FM, or sampled sound, with a hierarchy of control parameters directing the density, frequency range and temporal evolution of the synthesized sound textures. With sampled sound as a source, particularly rich textures may result from extremely small fragments of source material. Since 1989, I have also been developing a technique that stretches the sound, in a manner I call variable rate time shifting. The technique leaves the original pitch intact, or alternatively, transposes the sound in each grain by a different frequency ratio. The technique is similar to the time-shifting work reported by Jones and Parks [1988] except that our goal is to lengthen the sound, not shorten it. In addition, the technique is designed to work in real-time unlike computationally intensive methods used by the phase vocoder. Compositional experience using the technique has been particularly rewarding [Truax, 1990, 1992b], and we are currently implementing the technique on a microprocessor controlled board with the Motorola 65001 DSP chip and a 68000 controller [Truax & Bartoo, 1992].

2 Interpolating between Fixed and Continuous Sampling

The real-time program GSAMX implements a sampled sound instrument for granular synthesis where each grain consists of a short segment of sampled sound with specified duration and offset time from the beginning of the sound sample. The synthesis instrument consists of a bank of simple envelope generators with specifiable duration and delay (in ms) between successive envelopes. Each generator produces a three-part linear envelope whose attack and decay portions are a specifiable fraction of the event duration [Truax, 1988]. Additional variables include the start sample (or offset) and the range of this variable. Up to twenty simultaneous streams of voices of this synthesis instrument are possible with the DMX-1000. The instrument is controlled by a scheduler program on the host PDP Micro 11 where each grain is initiated and terminated under clock interrupts set at 1 ms. The shorter the grain duration, the higher the overall density of grains per second (gps). The minimum grain duration that can be effectively controlled in real-time is 10 ms, hence densities of up to 2000 gps can be achieved with the 20 simultaneous voices. However, at lower densities, grain durations may be as short as 2 ms.

Because each grain has an attack and decay, there is no possibility of clicks or transients depending on the portion of sampled sound being used. Moreover, when the grain streams are unarchived because each grain has a different duration or delay time between grains, and when each grain starts at a different position within the sound sample, very complex textures can result from even a very simple source sound.
Initially, two contrasting approaches were developed within the program as to the treatment of the sampled sound, one using a fixed block of samples, the other accepting continuous sample input from disk. Both with these two approaches suggested the need for a method which would interpolate between them, that is, to vary the rate at which new samples were introduced into the granulation process. I call this a "variable rate" approach. The desired effect is to be able to depart from the normal time flow of the continuous sampling model in a manner which would eventually approach the "frozen" time of the fixed sample model. Such an interpolation preserves the ongoing development of forward time flow but combines it with the sense of magnification of the moment associated with the fixed version.

3 Variable Rate Sampled Sound

In the variable rate implementation, the key variable is the rate at which new sound samples enter the DMSX's memory from disk compared with the synthesized output. The "rate" of the time-shifted sound is defined as the ratio of "off" milliseconds to "on" milliseconds and is called the "off/on ratio". Therefore a ratio of 0:1 is normal speed because there is no "off" time, and a ratio of 9:1 results in 9 ms of no forward movement through the sample before there is 1 ms shift forward, thereby producing a hundredfold time extension of the sample. However, since the grains are always taken from the current memory at one sample per calculated output, the frequency of the source material is not distorted, only the rate at which one advances through it in a macro-level sense. That is, micro-level waveform patterns and macro-level temporal changes have been effectively separated. The amount of time shift, described as the Time Extension Factor (TEF), can be calculated from the off/on ratio as follows:

\[ \text{Time Extension Factor} = \frac{(\text{Off Ratio} + \text{On Ratio})}{\text{(On Ratio)}} \]

(1)

Therefore, the ratio 1:1 produces a TEF of 2 times normal speed, and a ratio of 99:1 produces a TEF of 1000 times normal speed. This latter example means that one second of sound can last over 16 minutes! Since the TEF is proportional to the ratio, there is a strong intuitive relationship between them. One advantage of this approach is that there is no limit, other than one's patience to listen, to the amount of time stretching. The on/off ratio may be typed in and stored as part of a preset, that is, a set of control variables that may be recalled with a single keystroke. Each component of the ratio may also be separately "synched-origined" to allow the ratio or its components to be ramped, thereby continuously changing the TEF.

3.1 Automated Rate Control

Two types of automated control of the rate are also available. The first temporarily reverts to real-time when the maximum sample amplitude in a given disk block falls below a user-specified threshold value, thereby eliminating lengthy pauses. The second automated control correlates the rate to the maximum sample amplitude, thereby slowing down higher amplitude sounds and speeding up lower amplitude sections. The amount of rate variation depends on the maximum rate value which the user selects. This maximum value is implemented during the blocks with peak amplitude, with proportionately smaller values used during blocks with other amplitudes.

Depending on the length of the time window over which amplitude is assessed in order to perform the correlation, attack transients may be smeared by being part of the highest amplitude portion of the sound. Since the effect of longer time stretching is to lose the temporal character of the sound in order to enhance its spectral makeup, this loss of transients may be compositionally limiting. Therefore, a simple modification of the amplitude correlation that effects it by one block or time window is useful. If the attack is preceded by relative silence, it will be given a minimum stretch, whereas the steady state will be stretched by the maximum amount. However, manual control of the rate via the presets has also proved to be extremely effective, even with rapid material such as speech. With normal human reaction time, particular vowels or consonants can be elongated in the midst of a speech stream.
3.2 Harmonizing

Once an independence of pitch and duration was achieved with this technique, it seemed desirable to add simultaneous sample transposition. There are several standard approaches to this problem, the simplest one called harmonizing based on skipping an integer number of samples (e.g. taking every second sample for an octave up, every third sample for the third harmonic, and so on) or repeating samples (e.g. using each sample twice for an octave down). Implementation of the harmonizer approach would be simple if it were not for the technique used to realize the time stretching, namely the alternation between freezing the contents of memory and advancing through it. During the periods in which the memory contents are "frozen", a pointer goes backwards through the stored samples to obtain the material for the grains. However, when the current time position advances during the "on" times, the pointer increment must be higher in order to continue progressing backwards at the same rate and thus maintain the original pitch without discontinuities. For instance, with a current address marking the most recent sample received and a desired offset number of samples referencing a point in the past at which the grain is to start, the address of the next sample (sample address) during the "off" and "on" times is determined:

\[
\begin{align*}
\text{off mode:} & \quad \text{current address} = \text{current address} \\
& \quad \text{sample address} = \text{current address} - (\text{offset + 1}) \\
\text{on mode:} & \quad \text{current address} = \text{current address} + 1 \\
& \quad \text{sample address} = \text{current address} - (\text{offset + 2})
\end{align*}
\]

A simple harmonizing scheme to transpose the material to harmonic N generalizes the expression in brackets to \( (\text{offset + N}) \) for the off mode and \( (\text{offset + N + 1}) \) for the on mode. However, having only an upward harmonic transposition is a severe limitation, both in the absence of lower transpositions and the wide spacing of the first few harmonics. An alternative scheme implemented in GSAMX chooses a harmonic series based on a fundamental achieved by dividing the untransposed frequency by a factor \( F \). The case of \( F = 4 \) is particularly attractive since it allows a downward transposition of two octaves, such that the 4th harmonic is the original pitch, plus 4 transposition levels in the octave above the original (harmonics 5-8). The equations that determine the sample address are modified as follows to achieve harmonic \( N \):

\[
\begin{align*}
\text{off mode:} & \quad \text{sample address} = \text{current address} - \left( F \times \text{offset + N} \right)/F \\
\text{on mode:} & \quad \text{sample address} = \text{current address} - \left( F \times \text{offset + N + 1} \right)/F
\end{align*}
\]

Note that where \( F \) is a constant, \( N \) can vary with each grain or stream of grains, thereby allowing simultaneous transposition to different pitch levels in a multiple voice implementation. The DMX version of this algorithm has a maximum of 15 simultaneous voices, each with its own transposition level and choice of stereo output channel. The results range from the expected chordal enhancement of the material with lower pitched material to interesting timbral enrichment with unpitched or high frequency material.

4 Psychoacoustic Implications

Gabor [1947] described the microphonic level of the grain as a quantum of sound whose parameters of frequency and time form a unit rectangle. If one "squeezes" the rectangle the time is the domain, that is, shortens the grain duration, the frequency domain expands in compensation, that is, the bandwidth increases. For instance, in spectral analysis, longer time windows are required to define the low frequency components accurately, and vice versa. By linking frequency and time at the macro level, granularity makes it possible to treat the two variables independently at the macro level as described here. However, at the macro level the perceptual results of time stretching work on a similar inverse relationship: as a sound is progressively stretched, the less one is aware of its temporal envelope and the more one experiences its timbral character. Ironically, with extreme stretching in time, a spectrum can be experienced psychoacoustically in the classical Fourier manner, namely as the sum of its spectral components. Brief acoustic events tend to be recognized non-analytically according to their overall loudness, as well as their temporal and spectral envelopes which are perceived as a gestalt pattern. With stretched sounds, one has time to refocus one's attention on the inner
spectral character of a sound, which with acoustic sounds is amazingly complex. Therefore, transitions from the original to the stretched versions provide an interesting shift from one dominant percept to another.

Two related phenomena are commonly experienced with time shifting, one being the emergence of momentary resonances that are often quite vocal in character, the other being the perception of increased volume, as distinct from mere loudness. Elsewhere I have described the first phenomenon as the emergence of “inner voices” [Truxx, 1992b] which suggested to me archetypal imagery that informed my works Pacific and Domination. Particularly surprising was the discovery of these voices, resembling a distant choir singing vowels, in the sound of ocean waves. My explanation of the effect is that momentary resonances that normally are too fleeting and non-repetitive to be identified become audible by being prolonged and reinforced with multiple overlaps. In general, time stretching is a unique way to bring out the inner complexity of a sound.

To explain the perception of increased volume in the sound, I went back to the gestalt concept of volume as “the perceived magnitude of a sound” which tends to increase with spectral richness (or resonance), reverberation, duration, and of course intensity. This concept was current in early psychoacoustics, and can be found as late as 1967 in [Olson, 1967, p. 266]. In an effort to stimulate renewed interest in a complex, multi-parameter concept such as volume, I have proposed [Truxx, 1992a] a working model whose dimensions are spectral richness, time, and “temporal density” which refers to the temporal spacing of independent spectral components such as multiple sound sources and phase-shifted or time-delayed events. Time stretching contributes to all three dimensions, hence the perception of greatly increased volume. The overlay of simultaneous grains enhances spectral richness, the lack of synchronization between simultaneous grain streams adds to the temporal density, and finally, the extended duration occurs along the time axis.

5 Conclusion

The complexity and dynamic quality of granulated sampled sound makes it an attractive alternative to methods based on looping and transposition. Moreover, the basic unit or ‘quantum’ of the grain is a more flexible building block for treating sampled sound, particularly because the grain envelope avoids transient clicks when extracting and combining arbitrary sample segments. When granular synthesis is used to produce time-extended textures, it has no resemblance to instrumental and other note-based music; instead, the acoustic result often brings out the inner character of environmental sound. However it is used, granular synthesis creates a unique sound world and suggests new ways in which the music made with it can be related to the external world.

References