Time Scale Modification Using a Sines+Transients+Noise Signal Model

Tony S. Verma and Teresa H.Y. Meng
Department of Electrical Engineering, Computer Systems Laboratory, Stanford University
verma@furthur.stanford.edu, http://www.stanford.edu/~darkstar

Abstract

We propose a method for the time scaling of digitally sampled audio signals using a three part signal model consisting of sines+transients+noise. The three part model provides an accurate and flexible parametric representation for a wide range of audio signals. Because the proposed time scaling method manipulates each of the model components separately, the method allows modified tonal components of the signal to preserve pitch, transient components of the signal to preserve edges and noise-like components of the signal to remain noisy. The method therefore provides robust and natural sounding time scale modifications for a large variety of signals.

1 Introduction

For robust and meaningful modifications on audio signals it is important the underlying model represents the signal under question in a relevant or coherent manner. A wide range of audio signals are intuitively composed of the following parts: tones, transients and noise; furthermore each of these components must be manipulated separately for meaningful time scale modifications. It is desirable, when time scaling a signal, for tonal components of a signal to speed-up or slow-down in time while preserving their original pitch. Simultaneously, it is desirable for edges of a signal move to the proper onset locations with respect to the new time scale but preserve their original duration. In addition, any underlying noise-like components in the signal should also speed-up or slow-down along with the tonal components but maintain their noise-like quality. To this end, we present a robust time scaling method based on a flexible three part signal model which splits a signal into the components of sines+transients+noise (S+T+N) [1, 2]. The S+T+N model is an extension of Spectral Modeling Synthesis (SMS) [3] that includes an explicit and flexible model for transient signals.

Although many methods for time scale modifications exist, most methods do not represent transients coherently and therefore can not modify transients in a meaningful way. Time domain techniques based on overlap-add methods lack a signal model and often impart undesirable artifacts on the modified signal. While phase vocoder and sinusoidal modeling methods represent slowly varying tonal signals coherently, these methods will not time scale transients and noise in a meaningful way because they lack an explicit modification model for these types of signals. While SMS time scales signals containing sines and noise meaningfully, transients once again break this model. Although the method for time scale modifications given in [4] does have an explicit model for sines, transients and noise, the S+T+N model is more flexible because transients are modeled in a parametric form, whereas in [4] transients are kept as samples in the time domain.

The next section of the paper reviews the S+T+N model. Section 3 describes the necessary parameter modifications on each part of the S+T+N model to give natural sounding time scale modifications. The final section demonstrates the time scaling method on an audio signal.

2 The S+T+N Model

The S+T+N model uses three explicit analysis/synthesis techniques to represent and allow modifications on an audio signal. Figure 1 shows the analysis block diagram while figure 2 shows the synthesis block diagram. The first step in the model is sinusoidal modeling. Although use of any of the many sinusoidal models, for example [5, 3, 6], is possible, we use the sinusoidal model described in [2] because it allows a tight coupling between the three portions of the S+T+N model. The sinusoidal model finds then

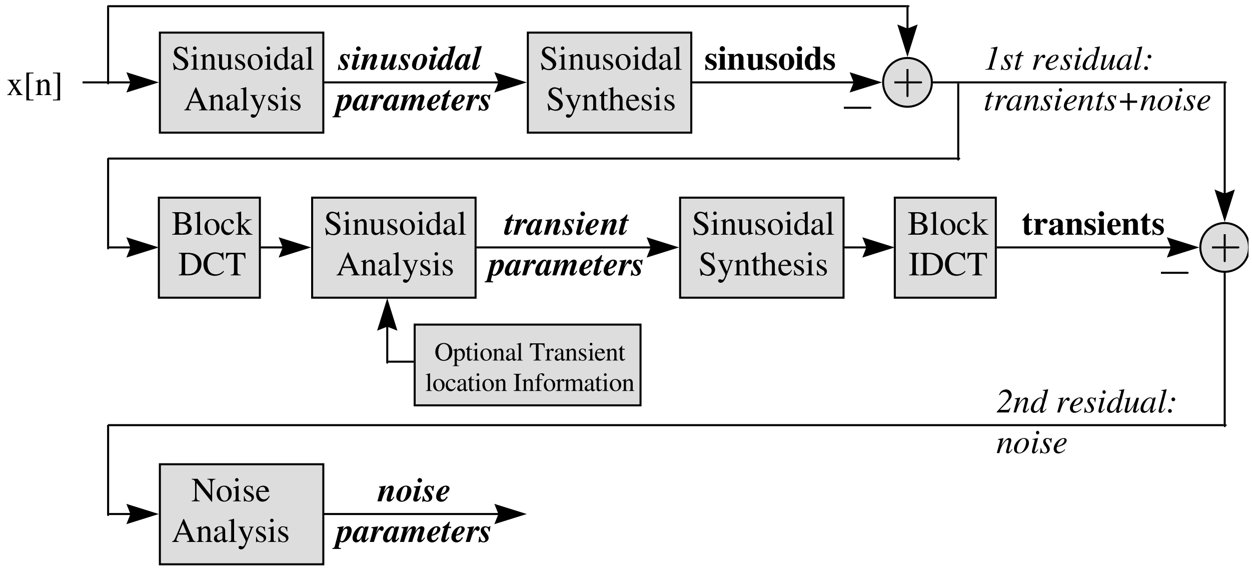

Figure 1: Analysis block diagram.

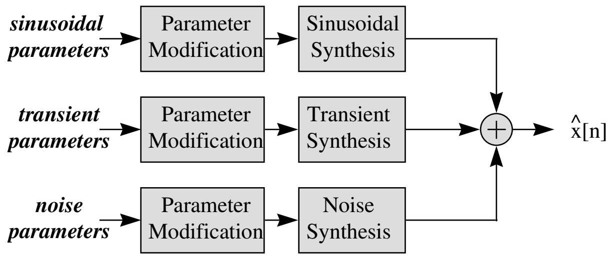

Figure 2: Synthesis block diagram.

removes the meaningful sinusoids in the signal creating a residual that consists of transients and noise. The transient model receives this residual, finds and removes the transient components from the residual leaving a second residual that contains noise, then passes the second residual to the noise model. Finally the noise model represents the second residual as a filtered random process using techniques such as in [3].

Because sinusoidal and noise models are widely used, we will not expand on them here. However the transient model from [1] is briefly reviewed in order for section 3 to make sense.

Although the S+T+N model has three parts, the analysis block diagram shows only sinusoidal and noise modeling blocks. Because of the duality between well developed sinusoids and transients, we can describe both sinusoids and transients using sinusoidal modeling provided we view the signal under question properly. This duality becomes apparent when observing the nature of these signals in the time and frequency domains. A slowly varying sinusoidal signal is impulsive in the frequency domain. This is why sinusoidal modeling is so effective at modeling slowly varying sinewaves. By performing a Short-Time Fourier Transform (STFT) analysis on the time-domain signal and tracking spectral peaks (the tips of the impulsive signals) over time, we can easily model slowly varying sinewaves. In contrast, transients, which are impulsive in the time-domain, cannot be easily tracked this way because its STFT analysis will not contain meaningful peaks. However, due to the duality between time and frequency, if transients are impulsive in the time-domain, they must be oscillatory in the frequency domain. Therefore we can track transients by performing sinusoidal modeling in a properly chosen frequency domain. The first step in the transient model is to map transient signals in the time domain to sinusoidal signals in some frequency domain. The Discrete Cosine Transform (DCT) provides such a mapping. It is defined as:

$$C(k) = \beta(k) \sum_{n=0}^{N-1} x(n) \cos\left[\frac{(2n+1)k\pi}{2N}\right]$$

for $n, k \in 0, 1, \ldots, N-1$ and $\beta = \sqrt{1/N}$ for $k = 1$, $\beta = \sqrt{2/N}$ otherwise. Roughly speaking, an impulse that occurs toward the beginning of a frame results in a DCT domain signal that is a relatively low frequency cosine. If the impulse occurs toward the end of the frame, then the DCT of the signal is a relatively

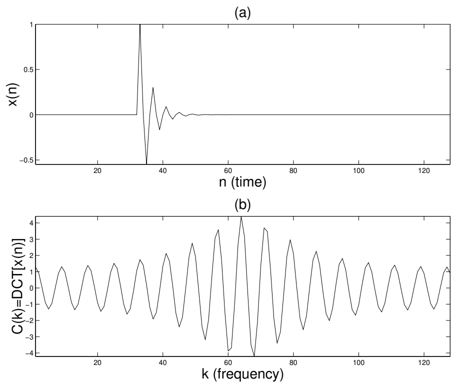

Figure 3: (a) Exponentially decaying sinusoid. A difficult signal for sinusoidal modeling. (b) DCT of exponentially decaying sinusoid. An ideal signal for sinusoidal modeling

high frequency cosine. This implies that changing the *frequency* of a DCT domain sinusoid will move the corresponding time domain impulse. This is the basis of moving transients without changing their duration which will be discussed in more detail in the next section.

Figure 3(a) shows a simple transient which is a one sided exponentially decaying sinewave. Performing sinusoidal modeling on this signal would be difficult for many reasons including meaningful parameter estimation and the number of sinusoids required to represent such an impulsive signal. Figure 3(b) shows the DCT of the transient signal. In contrast to the time-domain signal, the DCT domain signal is exactly the type of signal that sinusoidal modeling performs best on; it is a slowly varying sinusoidal wave. Therefore, by performing sinusoidal modeling in the DCT domain, we are actually modeling time-domain transients.

The previous discussion leads to a simple algorithm for an effective analysis/synthesis transient modeling tool. During the analysis, take non-overlapping blocks of the input signal. On each block perform a DCT. Now perform sinusoidal modeling on each DCT domain signal. This will result in model parameters that correspond to time-domain transients. Synthesis of the transients involves reconstructing the DCT domain sinusoids then taking an Inverse Discrete Cosine Transform (IDCT) to finally reconstruct the time-domain transients. Because the transient parameters are actually in a time-like domain, a 'fast' transient reconstruction that avoids the use of the IDCT is possible [7].

3 Time Scale Modifications

The modifications required for the sines and noise part of the model follow the discussion in [3]. During synthesis, use a different set of points in time than the analysis. That is, both the sines and noise part of the model use a particular hop-size during analysis. Increase/decrease the hop-size during synthesis for slowing down/speeding up the time progression of the sound. Transient modification also requires a time-scale modification, although in the DCT domain, in order to move transients to their proper onset locations without stretching or contracting them. At first, because frequency information of the transient model corresponds to time location, it would seem the necessary modification to the transients would be to multiply the transient frequency information by some factor (which is equivalent to pitch shifting the DCT domain sinewaves) in each DCT block. This, however, only translates the transients within each DCT block. To get the transients to match up with the new time-scale, we need to change the DCT block length while simultaneously translating the transients. The operation that does this is to modify the time-scale of the DCT domain sinusoids (i.e., time-scaling in the DCT domain) by the same factor as the time-domain sines and noise. This insures the reconstructed transient signal is the proper length (same number of samples) for summation with the modified sines and noise parts of the model. In addition, time expansion/compression of the DCT domain sinewaves causes more/fewer cycles of each sinewave to appear in the DCT block which effectively increases/decreases the frequencies in the DCT domain. This causes the transients to move to their proper onset location.

4 Example

As an example, we show the S+T+N time scale modification on a digital audio signal sampled at $f_s = 22KHz$; the results are shown in figure 4(a)-(f). Figure 4(a) shows the input signal, an excerpt from a drum solo with tom-tom hits. Figure 4(b) shows the input signal time stretched by factor of 1.5. Figures 4(c)-(f) show more detail of the time stretch by revealing the modifications of the sines and transients portion of the S+T+N model. Figure 4(c) shows the tonal portion (the perceived pitch of the tom drum) of the input signal extracted by the sinusoidal model, whereas figure 4(b) shows the time stretched tones. Although the stretched tones are 1.5 times longer in duration than the original tones, the pitch of the tones remains the same. Figure 4(e)

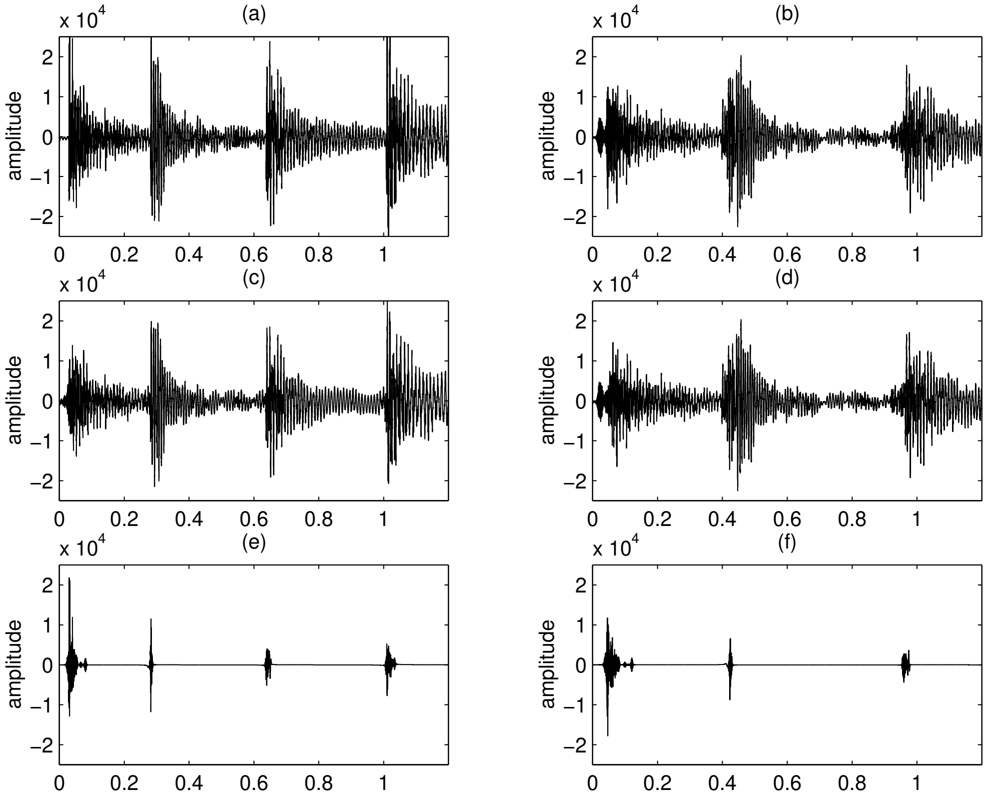

Figure 4: (a) Input signal (b) Time scaled S+T+N signal (c) Sines. (d) Time scaled Sines. (e) Transients. (f) Time Scaled Transients.

shows the attacks of the input signal extracted by the transient model, whereas figure 4(f) shows the time stretched transients. Although the time scaled attacks properly align with the stretched tones, the individual attacks are not stretched. Although not shown, the noise stretches with the tones but remains noise-like. Therefore the time stretched signal in figure 4(b) has been modified in a meaningful way.

References

[1] T. Verma, S. Levine, and T. Meng, "Transient modeling synthesis: a flexible transient analysis/synthesis tool for transient signals", in *Proc. ICMC*, September 1997, pp. 164–167.

[2] T. Verma and T. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio", in *Proc. ICASSP*, May 1998.

[3] X. Serra and J. O. Smith, "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition", *ICMJ*, vol. 14, no. 4, pp. 14–24, WINTER 1990.

[4] K. N. Hamdy, A. H. Tewfik, T. Chen, and S. Takagi, "Time-scale modification of audio signals with combined harmonic and wavelet representations", in *Proc. ICASSP*, April 1997.

[5] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal speech model", *IEEE Trans. ASSP*, pp. 744–754, 1986.

[6] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones", *JAES*, vol. 40, no. 6, pp. 497–515, June 1992.

[7] T. Verma and T. Meng, "A flexible sines+transients+noise model for audio", *To be submitted to ICMJ.*