AN ANALYSIS/SYNTHESIS TOOL FOR TRANSIENT SIGNALS THAT ALLOWS A FLEXIBLE SINES+TRANSIENTS+NOISE MODEL FOR AUDIO

Tony S. Verma and Teresa H.Y. Meng

Department of Electrical Engineering, Center for Integrated Systems
Stanford University
verma@furthur.stanford.edu, http://www-leland.stanford.edu/~darkstar

ABSTRACT

We present a flexible analysis/synthesis tool for transient signals that extends current sinusoidal and sines+noise models for audio to sines+transients+noise. The explicit handling of transients provides a more realistic and robust signal model. Because the transient model presented is the frequency domain dual to sinusoidal modeling, it has similar flexibility and allows for a wide range of transformations on the parameterized signal. In addition, due to this duality, a major portion of the transient model is sinusoidal modeling performed in a frequency domain. In order to make the transient and sinusoidal models work more effectively together, we present a formulation of sinusoidal modeling (and therefore transient modeling) in terms of matching pursuits and overlap-add synthesis. This formulation provides a tight coupling between the sines+transients+noise model because it allows a simple heuristic, based on tonality, as to when an audio signal should be modeled as sines and/or transients and/or noise.

1. INTRODUCTION

Sinusoidal modeling has enjoyed a rich history in both speech and audio [1, 2, 3]. One of the goals behind sinusoidal modeling is to allow meaningful transformations on the signal including time and pitch modifications. One problem with sinusoidal modeling is its difficulties in handling transient signals [4, 5, 6]. Because of these difficulties, transformations on transient signals lack meaning and flexibility. In [5], a flexible analysis/synthesis tool for transients signals was introduced. Here we present extensions and refinement to that work. The driving goal behind the model presented is to provide a flexible low order model for transient signals that fits well with current sinusoidal and sines+noise models. Extending these models to sines+transients+noise provides a more robust signal model and is essential for synthesizing realistic attacks of many instruments. The first section of the paper examines the need for an explicit transient model. The next section describes the flexible model for transient signals in terms of a duality to sinusoidal modeling. Because of this duality, the transient model is closely related to sinusoidal modeling. We describe a novel formulation for sinusoidal modeling, and therefore transient modeling, in terms of matching pursuits and overlap-add synthesis in section 4. The next section discusses how the sines+transients+noise model fits together. In addition, we describe how the matching pursuit/overlap-add formulation provides a tight coupling between the model's components because it allows a simple heuristic, based on tonality, as to when an audio signal should be modeled as sines and/or transients and/or noise. The final section gives an example.

2. MOTIVATION FOR A TRANSIENT MODEL

McAulay and Quatieri's Sinusoidal Transformation System (STS) [1] and Serra and Smith's Spectral Modeling Synthesis (SMS) [2] find sinusoidal components in a signal by spectral peak picking algorithms. Subtracting the synthesized sinusoidal components from the original signal creates a residual that consists of components that are not well modeled by sinusoids. These components are transients and noise [3, 4]. Although transients and noise can be modeled by a sum of sinusoidal signals, as in the case of the Fourier Transform, it is not necessarily efficient or meaningful. In the STS system, generally applied to speech, the transient+noise residual is often masked sufficiently to be ignored [1]. In audio applications, this residual is important to the integrity of the signal. The SMS system furthers the sinusoidal model by explicitly modeling the residual as slowly filtered white noise. Although this technique has had success, transients do not fit well into this model either. Transients modeled as filtered noise lose sharpness in their attack and sound dull. As suggested by others [2, 6, 7], transients need to be handled separately from both noise and sinusoids. One method that has been considered is removing transient areas from the residual, performing noise analysis, then adding the transients back into the signal [4, 7]. Although this method works, it has drawbacks. First it lacks flexibility because there is no model for the transients; they are left in their original format. Secondly many instruments have an underlying noise, the breathiness of a flute for example, that is neither sinusoidal nor transient. Removing transient in the fashion stated, both transients and noise are removed. It is desirable to model transients separately but leave noise to the noise model. Another approach is to insure the sinusoidal model can handle any type of signal, including transients and noise, as is the case in [3]. For many transformations, however, it makes sense to transform the sinusoidal components differently from the transient components which is not possible if the sinusoidal model handles the entire signal. With an all inclusive sinusoidal model, it is not clear if transformations will be natural [8]. The need for an explicit flexible transient model that allows for transformations and fits well into current sines+noise models motivate the transient model presented here.

3. THE TRANSIENT MODEL

The underlying theme behind the transient model is that it is the frequency domain dual to sinusoidal modeling. Because of this duality, the parameters that characterize the sinusoidal components of a signal also characterize the transient components of a signal, al-

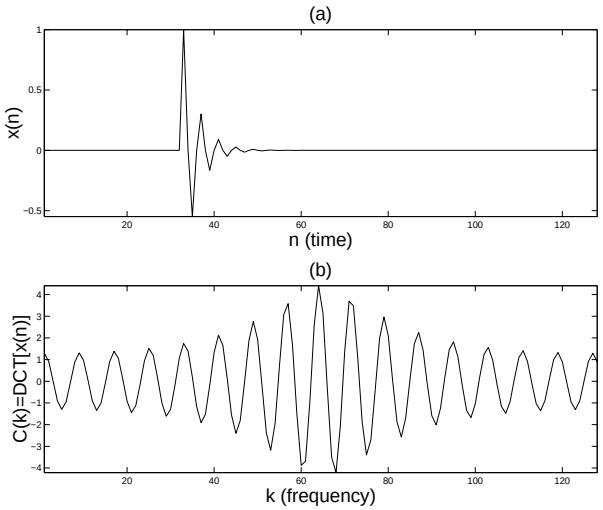

Figure 1: (a) Exponentially decaying sinusoid. A difficult signal for sinusoidal modeling. (b) DCT of exponentially decaying sinusoid. An ideal signal for sinusoidal modeling

though, as will be shown, in a different domain. Additionally, this allows the core components of the transient modeling algorithm to be identical to the sinusoidal modeling algorithm.

3.1. Duality between sines and transients

There is a isomorphic duality between well developed sinusoids and transients. That is we can describe both sinusoids and transients with the same tool, namely 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 $\beta = \sqrt{\frac{1}{N}}$ for $k = 1$, $\beta = \sqrt{\frac{2}{N}}$ 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 high frequency cosine. Transients encountered in real audio signals, however, are not generally ideal Kronecker Delta functions. Figure 1a shows a more realistic 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 1b 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 sinewave. Therefore by performing sinusoidal modeling in the DCT domain, we are actually modeling time-domain transients.

3.2. Algorithm for transient modeling

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. This will result in model parameters that correspond to time-domain transients. The combination of the DCT then STFT analysis to find meaningful peaks takes the signal from the time-domain into the DCT frequency domain and then back into some type of time-like domain. Although it may seem redundant for the transient model to perform these transformations, theses operations rotate (unitary transforms simply rotate vector spaces) the signal in such a way to make transients readily apparent. 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.

4. MATCHING PURSUIT/OVERLAP-ADD FORMULATION

A major portion of the transient model, because of the duality previously discussed, is sinusoidal modeling. Many methods for sinusoidal modeling exist [1, 2, 3]. Here we present a new formulation of sinusoidal modeling, and therefore a major part of the transient model, as a variation of matching pursuits with overlap-add synthesis.

4.1. Matching pursuits

Matching pursuits refers to an iterative method for computing signal decompositions in terms of a linear combination of vectors from a highly redundant dictionary [9]. The M elements of the dictionary, $\mathcal{D} = \{g_m\}; m = 1, 2, \ldots, M$, span $\mathcal{R}^N$ and are restricted to have unit norm, $\|g_m\| = 1$ for all m. The algorithm is greedy in that at each stage the vector in the dictionary that best matches the current signal is found and subtracted to form a residual. The algorithm then continues on this residual signal. More specifically, in the first stage of the algorithm, the first residual is set equal to the input signal: $r_1 = x$. The first index, m_1^*, that corresponds to the dictionary element that has the largest correlation with the first residual is found. This index maximizes $|\left\langle g_{m_1^*}, r_1 \right\rangle|$ over all m. The projection onto this dictionary element is then subtracted from the current residual to form the next stage residual. Thus at the k^{th} iteration, for $k > 1$, the residual signal is $r_k = r_{k-1} - \alpha_{k-1} g_{m_{k-1}^*}$, where $\alpha_{k-1} =< g_{m_{k-1}^*}, r_{k-1} >$. Therefore, the decomposition consists of a set of weighting terms

$\{\alpha_1, \alpha_2, \ldots\}$ and indices $\{m_1^*, m_2^*, \ldots\}$. The signal reconstruction is the linear combination of the dictionary elements found during the decomposition. If the decomposition ran for K iterations, then the reconstruction is: $\sum_{k=1}^{K} \alpha_k g_{m_k^*}$.

The energy in the residual converges to zero as the number of iterations approaches infinity [9]. Although exact reconstruction is possible, the matching pursuit is generally stopped by some criteria to allow low order approximations to the input signal. For the purposes of a sines+transients+noise model of audio, the stopping criteria is important and will be considered in more detail in section 5.

4.2. Overlap-add formulation

In our formulation of sinusoidal modeling, frames of the signal x are represented as a combination of sinusoidal signals. The combination is found via matching pursuits. These frames are then combined in an overlap-add fashion to reconstruct the entire signal. Mathematically we must take $x : \{x[n]; n \in \mathcal{Z}\}$ constrained to be in $l_2(\mathcal{Z})$, and make an ensemble of timelimited signals $x_l : \{x_l[n]; l, n \in \mathcal{Z}\}$ by hopping a rectangular window over signal. Define the lth windowed signal as

$$x_l[n] = \sqcap_N \left[n - l\frac{N}{2} \right] x[n] \tag{1}$$

Where a hopsize of half the window length is assumed and

$$\sqcap_N[n] = \begin{cases} 1 & n = 0, 1, \ldots, N-1 \\ 0 & \text{otherwise} \end{cases}$$

is the rectangular window. Each of these timelimited signals can then be considered a finite duration signal in $\mathcal{R}^N$ to which matching pursuits are applied. Although we consider each timelimited signal a finite duration signal in order to apply matching pursuits, we keep track of the time location of each frame to ensure proper reconstruction. Therefore the matching pursuit reconstruction of each frame, $\hat{x}_l : \{\hat{x}_l[n]; l, n \in \mathcal{Z}\}$, is once again considered an ensemble of timelimited signals. Finally, the approximation to x, $\hat{x} : \{\hat{x}[n]; n \in \mathcal{Z}\}$, is completed with a windowed overlap-add reconstruction of the form:

$$\hat{x}[n] = \sum_l w \left[n - l\frac{N}{2} \right] \hat{x}_l[n] \tag{2}$$

Where the reconstruction window w is a timelimited function with the constraint $\sum_l w \left[n - l\frac{N}{2} \right] = 1$. If the error of the matching pursuit on each windowed signal is allowed to converge to zero, the formulation yields a perfect reconstruction system which is immediate from plugging (1), which is the error-free matching pursuit decomposition of each frame, into equation (2).

4.3. The matching pursuits dictionary

It still remains to define the dictionary to use for matching pursuits. We could use cosines indexed by parameters of uniformly spaced $frequency$ and $phase$. The $amplitude$ parameter would be found by the correlation computation. This formulation is possible, but the dictionary is parameterized by both $frequency$ and $phase$ which leads to inefficiency in the computation of the matching pursuit. By using a generalization of the matching pursuit algorithm developed in [10], the phase parameters can be found as part of the correlation computations. In addition, the correlation computations can be efficiently computed by use of the Fast Fourier Transform (FFT).

The generalizations in [10] allow each iteration of the matching pursuit to find optimal dictionary subspaces as opposed to finding the optimal dictionary element. Choosing the subspace as a dictionary element and its complex conjugate allows many simplifications in the computations required for the matching pursuit formulation. We now consider a dictionary that consists of complex exponentials, $g_m(n) = e^{j2\pi f_m n}$ and its complex conjugates. This dictionary is indexed only by the $frequency$ parameter. As shown in [10], if the signal to be decomposed is real, then the expansion coefficients appear in conjugate pairs and the new residual at each stage of the matching pursuit is also real. In addition, because the choice of dictionary elements are complex exponentials, at each iteration, the projection onto the dictionary subspace at will be a constant amplitude, constant frequency cosine. The $amplitude$ and $phase$ for each of the cosines are found from the set of weights $\{\alpha_1, \alpha_2, \ldots\}$, i.e., from the correlation computations, and the $frequency$ for each is found from the set of indices $\{m_1^*, m_2^*, \ldots\}$.

Since each iteration of the matching pursuit requires M correlation calculations, after which the largest absolute correlation must be found, the computational complexity is high. The computational burden can be lessened by updating the correlations at each iteration using [9]:

$$\langle g_m, r_k \rangle = \langle g_m, r_{k-1} \rangle - \alpha_{k-1} \left\langle g_m, g_{m_{k-1}^*} \right\rangle, \text{ for all } m \tag{3}$$

Thus for any application of matching pursuits only one set of M correlations need be computed at the start and the rest of the correlations for subsequent residual signals can be updated iteratively. In addition, because our dictionary for sinusoidal modeling consists of uniformly spaced complex exponentials and their complex conjugates, we can use the Discrete Fourier Transform (DFT) (or the FFT if the number of dictionary elements are a power of 2) for the initial correlation computation. The amount of zero-padding in the DFT computation determines the redundancy of the dictionary. Furthermore, at each iteration, equation (3) in terms of the DFT, says that the windowed (in our formulation, a rectangular window) transform of the previous iterations projection should be subtracted from the previous iteration's correlations (or DFT) to get the updated correlations (or DFT).

5. THE COMPLETE MODEL

With the transient model and the matching pursuits/overlap-add formulation of sinusoidal modeling described, a flexible sines+transient+noise model for audio signals is now presented. The first step in the model is sinusoidal modeling. The meaningful sinusoids are modeled and removed from the signal, creating a residual signal, using the sinusoidal model formulation previously described. This residual signal that now consists of transients+noise is processed by the transient model. Transients are modeled and removed by the techniques previously described. Finally, the signal that now consists of noise is modeled as a filtered random process using techniques in the literature [2, 6].

Formulating sinusoidal modeling in terms of matching pursuits and overlap-add synthesis leads to an algorithm that is similar to the analysis-by-synthesis sinusoidal modeling algorithm in

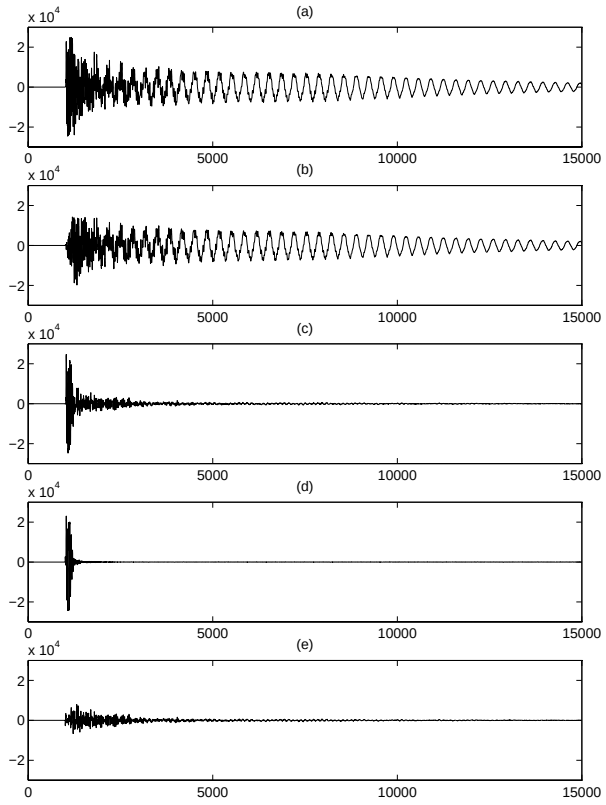

Figure 2: (a) Original xylophone. (b) Synthesized sinusoids. (c) First residual containing transients+noise. (d) Synthesized transients. (e) Second residual containing noise

[3]. An important question in a matching pursuit or analysis-by-synthesis is when to stop the algorithm. In [3], because only sinusoidal components are used in the model, the algorithm continues until the residual is acceptably small. If the signal is transient or noise-like the algorithm will require many iterations resulting in many sinusoidal parameters. In our formulation, because we use explicit sinusoidal, transient and noise models, we need to stop the sinusoidal algorithm before transients or noise are modeled as sinusoids. In addition, we need to stop the transient model matching pursuit before noise is modeled as transients.

To this end we use a heuristic based on the tonality of the signal. The tonality is measured as in [11]. Using this measure, the matching pursuits iterations continue until the residual is no longer tonal. This works for both the sine and transient portions of the model because in the time-domain if a signal is not tonal it does not contain time-domain sinewaves and in the DCT domain if signal is not tonal it does not contain time-domain transients.

6. EXAMPLE

As an example, we show the sines+transients+noise analysis on a xylophone hit, the results of which are shown in figure 2. The xylophone, although inharmonic, has a perceived pitch which is modeled well by the sine portion of the representation. Figure 2(a) is a plot of the original signal sampled at $44.1KHz$, while figure 2(b) shows the synthesized sinusoids. Figure 2(c) is the first residual which shows the sharp attack of the sound as well as some underlying noise. The attack, as modeled by the transient model, is shown in figure 2(d). Figure 2(e) shows the second residual which is the part of the original signal that is not well modeled by sines or transients. This is slowly varying noise. If the first residual signal were passed to the noise model without the transient model, the attack would be smeared and the characteristic 'knock' of the xylophone would be lost. The summation of the sines+transient+noise portions yield a signal that is perceptually indistinguishable from the original.

7. CONCLUSION

Combining the transient model with current sines+noise models allows parameterization of a wide range of sounds while the synthesized versions are perceptually identical to the original. Because many sounds can be meaningfully thought of in terms of the components of sines, transients and noise, transformations that are intuitive and natural are possible.

8. REFERENCES

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

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

[3] E. George and M. 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, Jun. 1992.

[4] X. Serra, *A System For Sound Analysis Transformation Synthesis Based on a Deterministic Plus Stochastic Decomposition*, PhD thesis, Stanford University, 1989.

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

[6] M. Goodwin, "Residual modeling in music analysis/synthesis", in *Proc. ICASSP*, May 1996, pp. 1005–1008.

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

[8] E. George and M. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model", *IEEE Trans. ASSP*, vol. 5, no. 40, pp. 389–406, Sep. 1997.

[9] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Trans. SP*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[10] M. Goodwin, "Matching pursuit with damped sinusoids", in *Proc. ICASSP*, Apr. 1997, pp. 2037–2040.

[11] ISO/MPEG Committee, "Coding of moving pictures and associated audio for digital storage media at up to about 5 1.5mbit/s - part 3: Audio", *ISO/IEC 11172-3*.