A SWITCHED PARAMETRIC & TRANSFORM AUDIO CODER

Scott N. Levine[*] *and Julius O. Smith III*

Center for Computer Research in Music and Acoustics
Department of Electrical Engineering
Stanford University
scottl@ccrma.stanford.edu, http://www-ccrma.stanford.edu/~scottl

ABSTRACT

In this paper, we present a system of sines+transients+noise modeling techniques that dynamically switches between parametric representations and transform coding based representations. The sines and noise are represented by parametric models using multiresolution sinusoidal modeling and Bark-band noise modeling, respectively. The transients are modeled by short regions of transform coding. In addition, new methods are presented for selection and quantization of sinusoidal trajectories based on trajectory length and signal-to-masking thresholds. This system is useful for both low bitrate audio coding (20-40 kbps) and compressed-domain processing, such as time-scale modification.

1. INTRODUCTION

The goal of this paper is to present a new representation for audio signals that allows for low bitrate coding while still allowing for high quality compressed domain processing, such as time-scaling modifications. In the current MPEG-4 specifications, there are compression algorithms that allow for time and pitch modifications, but only at very low bitrates (2-16 kbps) and relatively low bandwidth (8 to 11 kHz sampling rate) using sinusoidal modeling or CELP [1]. In this system, we strive for higher quality with higher bitrates (20-40 kbps), while allowing for a larger audio bandwidth (32 kHz sampling rate) and high quality time scale modifications.

For very low bitrates, audio compression design becomes a problem of deciding what kinds of artifacts to introduce, and deciding how much to reduce them in exchange for a reduced audio bandwidth. This is due to the fact the neither transform, nor switched parametric/transform coding can be perceptually lossless at these bitrates. By utilizing parametric coding at these low bitrates, especially noise modeling, a larger audio bandwidth is attainable than by using just transform coding [2] alone.

To achieve the data compression rates and wideband modifications, we first segment the audio (in time and frequency) into three separate signals: a signal which models all sinusoidal content with a sum of time-varying sinusoids [3], a signal which models all attack transients present using transform coding, and a Bark-band noise signal [4]. Each of these three signals can be individually quantized using psychoacoustic principles pertaining to each representation.

Transform coding is used for the transients instead of sinusoidal modeling [5] or noise modeling [6, 4] because of their inability to accurately encode the waveform efficiently. Transient attacks of instruments can be very sudden and broadband, and these are notoriously difficult signals for parametric coders such as sinusoidal models. During a transient, transform coding is used to represent the signal. At all other times, sinusoidal and noise modeling represent the signal. Because of phase-matching algorithms, the parametric and transform systems can switch seamlessly.

High-quality time-scale modifications are now possible because the signal has been split into sines+transients+noise. The sines and noise are stretched/compressed with good results, and the transients can be time-translated while still maintaining their original temporal envelopes. In time-scaled (slowed) polyphonic music with percussion or drums, this results in slowed harmonic instruments and voice, with the drums still having *sharp* attacks.

In this paper, we will first describe the system from a high level point of view, showing how the input audio signal is segmented in time and frequency. We will then spend one section on each of the three signal models: sines, transients, and noise. In each of these sections, we will also describe their separate methods of parameter quantization. Afterwards, another section will be devoted to compressed-domain time-scale modifications.

2. SYSTEM OVERVIEW

This system segments the audio signal into sines, transients, and noise. The first segmentation performed is between transient and non-transient regions. During non-transient regions, the signal is represented with multiresolution sinusoidal modeling [3] between 0 and 5 kHz, which will be discussed in Section 3. The residual of the original signal minus the synthesized sinusoids is modeled by a variant of Bark-band noise modeling [4], to be detailed in Section 5. Therefore, between 0 and 5 kHz, the non-transient signal model consists of sines and noise. Between 5 and 16 kHz, there is only noise modeling. Sinusoidal modeling is not performed above 5 kHz to reduce the overall bitrate. During transient regions, which last approximately 70 msec, transform coding is performed, as described briefly in Section 4. Careful phase matching is performed during the transition between sines and the transients, so that no discontinuities are heard, even when time-scaled. More on this will be discussed in Section 3. For a graphical example, see Figure 1, which shows both the time-domain signal of a drum attack and its corresponding time-frequency plane.

The transient detector carefully decides when to switch between the parametric sines+noise model to the transform coded transient model. For added stability of the detector, it relies upon two different measures. The first measure looks for rising edges

[*]Work supported by Bitbop Laboratories

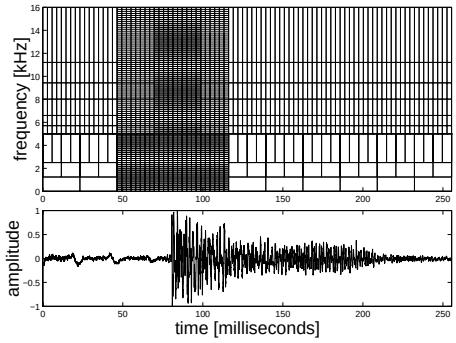

Figure 1: The initial time-frequency segmentation into sines, transients, and noise. During the time region surrounding the attack, transform coding is used. During the other times, multiresolution sinusoidal modeling is used below 5 kHz, and Bark-band noise modeling is used from 0 to 16 kHz.

in the short-time energy of the highpassed original input signal. The second measure examines the ratio of the sinusoidal modeling residual short-time energy to the original signal short-time energy. If the residual has a low relative energy, then the input signal is well modeled by the sinusoidal modeling (in a mean square sense). If the residual is as large as the original signal (or larger), then *pre-echo* artifacts are likely, and a sudden transient may be present that sinusoidal modeling could not well represent. By determining when both of these measures are concurrently above a given threshold, a transient region is determined.

3. MULTIRESOLUTION SINUSOIDAL MODELING

Sinusoidal modeling represents an audio signal by a sum of time-varying oscillators, whose amplitude, frequency, and (optionally) phase parameters are updated every frame [5, 6]. From frame to frame, the sinusoidal parameters are grouped into trajectories. The parameters in each trajectory are guaranteed to have a limited deviation in amplitude and frequency over time. The trajectories can have length as short as one frame, or as long R frames, where R may be set to bound the latency of the system.

When the sinusoids are to be synthesized at the decoder, the sinusoidal parameters are interpolated at a sample-rate from the frame-rate parameters in their corresponding trajectory. Instead of parameters updated at every frame boundary, now the parameters are updated every sample. This intra-frame interpolation can lead to timing inaccuracies in the synthesis. For example, when a sinusoid abruptly begins, the synthesized version has an amplitude that is linearly interpolated from zero in the frame preceding the onset to the full original amplitude in the frame containing (or following) the onset. Because the synthesized audio begins before the original signal, it is termed *pre-echo*. The shorter the parameter estimation window, the less pre-echo is a problem. Unfortunately, frequency resolution of the estimation process reduces as the window length reduces.

In order to reduce pre-echo artifacts, the signal is split into three multiresolution octave-spaced signals [3]. Parameter estimation is performed individually on each octave signal; the lowest octave has good frequency resolution, but has poor time resolution. The highest octave has good time resolution, but has poor frequency resolution. The reasoning behind this system is that due

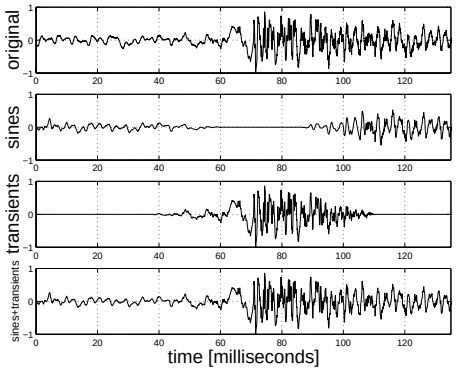

Figure 2: In the top plot, the original signal is shown containing speech and a bass drum hit at time=65 milliseconds. The second plot shows the synthesized multiresolution sinusoids, which are faded out during the transient. The third plot shows the transform-coded transient, which is the residual between the original and the sinusoids. Only during the frames that the sinusoids are faded in and out, cubic polynomial phase interpolation is used in order to guarantee phase locking with the transient. The bottom plot shows the sum of the sines and transients.

to the near-logarithmic perception of pitch, frequency resolution is more important at lower frequencies than at higher frequencies.

3.1. Sinusoidal Phases

It was mentioned in the previous section that sinusoidal modeling normally parameterizes amplitudes, frequencies, and phases. But for most music, phase information is not needed unless one is encoding a transient, or one needs to compute a residual. The noise is computed from the sinusoidal residual, but it is not perceptually important for sines and synthesized noise to be phase locked. Also, sinusoidal modeling is not utilized during transients; transform coding is used instead. Therefore, while sinusoidal modeling is representing steady-state tones, the phases of the sinusoids are let to unwind freely as long as certain frame boundary conditions are met. This is sometimes referred to as *phaseless* reconstruction. Phaseless reconstruction does not need any explicit transmitted phase information. But, there is a transition region between sinusoidal modeling regions and transients, where both sinusoids and transform coded data overlap; one is being faded out while the other is being faded in. During this transition, which can be seen in Figure 2, the phases of the sinusoids must be correctly aligned at the decoder so as to correctly match the phase of the transform coded data (computed from the sinusoidal residual). In order to assure this, explicit phase information is transmitted for the sinusoids just before and after the transient region. Cubic-polynomial phase interpolation [5] is used only during this region to assure correct phase alignment. For more details, see [7, 8].

3.2. Sinusoidal Trajectory Selection

With all the sinusoidal parameters estimated and formed into trajectories, the next issue is to decide which trajectories to keep and which to eliminate. Ideally, only trajectories modeling true sinusoids would be kept; any attempting to model noise processes would be eliminated, since these will be later picked up by the

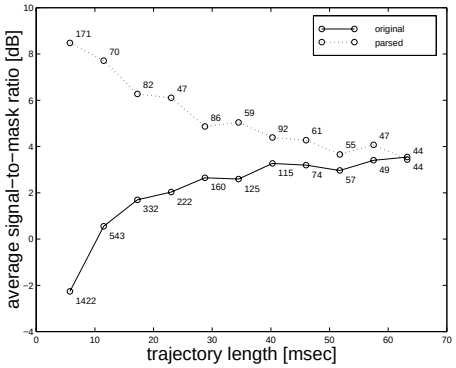

Figure 3: SMR statistics of trajectory length of the original parameter estimation (lower solid curve) and the results of the trajectory selection process (upper dotted curve). The numbers next to each circled point on the curves show the total number of trajectories at the given length in the analyzed audio signal.

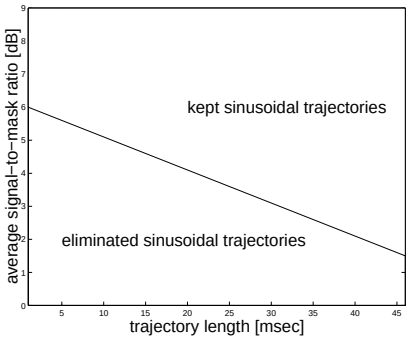

Figure 4: Any trajectory in the bottom left region of the average-SMR/trajectory-length space in the above figure will be eliminated.

Bark-band noise modeling algorithm. This selection process will lower the bitrate and improve the quality of time-scale modifications. The first step in the selection process is to compute a signal-to-masking ratio (SMR) threshold for the synthesized multiresolution sinusoids. Then, for each trajectory, a time-averaged SMR is computed. As can be seen in the lower solid curve in Figure 3, the original parameter estimation returns short trajectories with low or negative SMR. As the trajectories become longer, the average SMR slowly increases. In addition, the number of trajectories is by far the greatest at length 1, and then (approximately) exponentially decreases as the trajectories get longer. These data suggest that the parameter estimation is modeling noise with many sinusoidal trajectories of short length. To eliminate these, a selection metric is devised that is a function of both trajectory length and average SMR. Any trajectory with either sufficiently short trajectory length or low SMR, or some linear combination of both, will be eliminated. As a result of the trajectory elimination, the only remaining short trajectories have a high SMR, as seen in the dotted curve in Figure 3. It can be assumed that these trajectories are actually short-lived spectral peaks. By eliminating many short and low SMR trajectories, the sinusoidal bitrates decreased by as much as 50%.

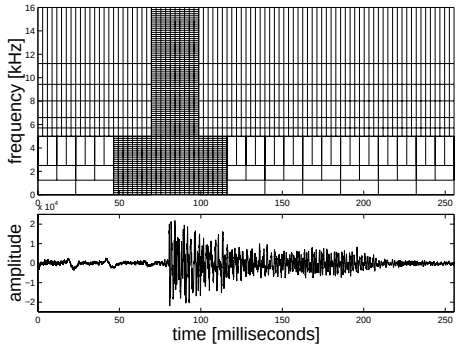

Figure 5: The time-frequency region of the MDCT coding of the transient is pruned in order to reduce the bitrate while still maintaining high perceptual quality.

3.3. Sinusoidal Trajectory Quantization

After the trajectory selection, each sinusoidal parameter is quantized to its just noticeable difference (JND) in amplitude and frequency. Any trajectory with a sufficiently low SMR is downsampled by two (in time), and then interpolated at the decoder. All trajectories are time-differentially encoded, scalar quantized, and Huffman encoded. For most signals, the sinusoids took 8 to 12 kbps.

4. TRANSFORM CODED TRANSIENTS

When the transient detector deems a given time region a transient, then that region is encoded using standard transform coding techniques [2]. Each window is 256 points long (at 44.1 kHz sampling rate), with 50% overlap, and is transformed using an MDCT. In total, 24 overlapping short windows are used across the transient region. Special care is taken at the transient region boundaries to assure that aliasing cancellation is provided for [8]. In order to reduce the overall bitrate, the time-width of the transient is reduced at high frequencies, from 70 msec to 30 msec. The time-frequency plane of this pruning procedure can be seen in Figure 5, which reduces the number of MDCT coefficients by 30%, with no audible distortions on the many pieces of music tested. The time-width of the transient MDCT coefficients can be modified individually for many frequency ranges, not just two as shown in Figure 5.

5. NOISE MODELING

Based on the work of [4], Bark-band noise modeling is used in the system. From 0 to 5 kHz, the residual between the original and the synthesized sinusoids is split into ten bands of equal width on the Bark scale (only during the non-transient regions) using FFT methods. For every 23 msec frame, an energy level is encoded for each band, and then quantized. At the decoder, a piecewise linear magnitude spectrum is formed using the band energy. Random spectral phases are used, and the noise is synthesized using an IFFT. On average, this residual requires 2 to 3 kbps. From 5 to 16 kHz, during non-transient regions, only noise modeling is present (no sinusoids are used). This region is split into six Bark bands, and the same noise synthesis procedure is performed on 3 msec frames. To reduce the bitrate, line segment approximation

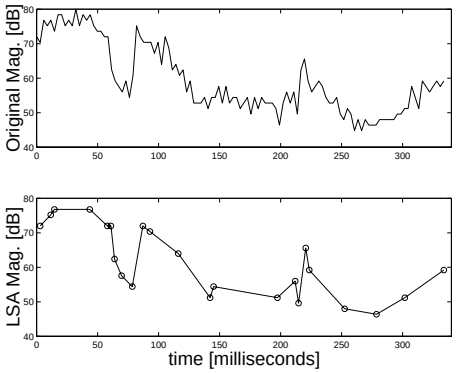

Figure 6: The top plot shows a Bark band (8000-9200 Hz) RMS-level energy envelope for about 300 milliseconds. The bottom plot shows the line segment approximated RMS-level energy envelope. The circled points are the transmitted envelope points, and the remaining points are linearly interpolated using the transmitted points.

is performed upon the energies in each band over time, as can be seen in Figure 6. This nonlinearly smoothes the trajectories, and represents the high frequency noise using 3 to 4 kbps.

6. COMPRESSED DOMAIN PROCESSING

One advantage of parametric coders is that compressed domain modifications are relatively simple to perform. However, compressed domain processing such as time-scale modification is also easy with a switched parametric/transform coder. Time-scaling the sines is a straightforward change to the synthesis frame length [6], and time-scaling noise is simply time-interpolating the Bark-band noise gains. In order to time-scale the transients, they are *not* stretched like the sines and noise. Rather, they are *translated* in time. This assures that all attacks will remain as sharp as they were in the original; they just now happen at different times. A graphic example of this process can be seen in Figure 7.

7. CONCLUSIONS

In this paper, we show that a low bitrate audio compression system can be designed with good quality and the ability to perform compressed domain processing. Both parametric and transform coders are used, and are dynamically and seamlessly switched depending on a transient detector.

8. REFERENCES

[1] B. Edler, "Current status of the MPEG-4 audio verification model development", *Audio Eng. Soc. Convention*, 1996.

[2] M. Bosi, K. Brandenburg, S. Quackenbush, L.Fielder, K. Akagiri, H.Fuchs, M.Dietz, J.Herre, G.Davidson, and Y.Oikawa, "ISO-IEC MPEG-2 Advanced Audio Coding", *Audio Eng. Soc. Convention*, 1996.

[3] S. Levine, T. Verma, and J.O. Smith, "Multiresolution sinusoidal modeling for wideband audio with modifications", *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Seattle*, 1998.

[4] M. Goodwin, "Residual modeling in music analysis-synthesis", *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Atlanta*, pp. 1005–1008, 1996.

[5] R. McAulay and T. Quatieri, "Speech transformations based on a sinusoidal representation", *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, December 1986.

[6] X. Serra, *A System for Sound Analysis/Transformation/Synthsis based on a Determistic plus Stochastic Decomposition*, PhD thesis, Stanford University, 1989.

[7] S. Levine and J. Smith, "A sines+transients+noise audio representation for data compression and time/pitch-scale modifications", *Audio Eng. Soc. Convention*, 1998.

[8] S. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, PhD thesis, Stanford University, expected December 1998, available online at http://www-ccrma.stanford.edu/~scottl.

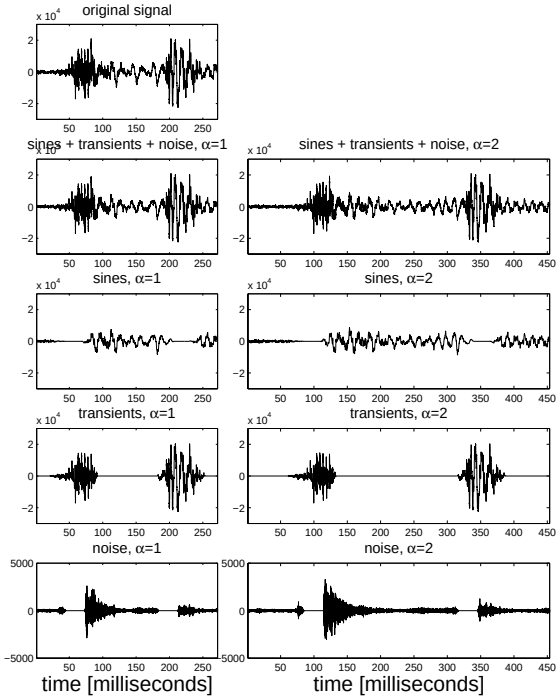

Figure 7: This set of plots shows how time-scale modification is performed. The original signal, shown at top left, shows two transients: first a hi-hat cymbal hit, and then a bass drum hit. There are also vocals present throughout the sample. The left-side plots show the full synthesized signal at top below the original, and then the sines, transients, and noise independently. They were all synthesized with no time-scale modification, at α=1. The right-side plots show the same synthesized signals, but time-scale modified with α=2, or twice as slow with the same pitch. Notice how the sines and noise are stretched, but the transients are only translated. Also, the vertical amplitude scale on the bottom noise plots are amplified 15 dB for better viewing.