Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

IMPROVEMENTS TO THE SWITCHED PARAMETRIC & TRANSFORM AUDIO CODER

Scott N. Levine

Liquid Audio
Redwood City, CA 94301
scottl@liquidaudio.com

Julius O. Smith III

Center for Computer Research in Music and Acoustics
Stanford University
jos@ccrma.stanford.edu

ABSTRACT

In this paper, we introduce improvements to previous sines + transients + noise audio modeling systems, including new sinusoidal trajectory selection and quantization procedures. In previous work [1], the audio is first segmented into transient and non-transient regions. The transient region is modeled using traditional transform coding techniques, while the non-transient regions are modeled using parametric sines plus noise modeling. Because such a system contains a mix of parametric and non-parametric techniques, compressed-domain processing such as time-scale modifications are possible.

1. INTRODUCTION

Sines + transients + noise systems attempt to model audio in such a manner that can achieve a competitive perceptual coding gain while allowing for high quality compressed domain modifications. Currently, the most efficient audio data compression systems are transform coding based [2], and inherently not parametric. Thus, transform coder based systems cannot perform compressed-domain modifications such as time and pitch scaling at a reasonable cost. Methods from the computer music world of parametric sines plus noise [3] modeling were optimized for high quality signal modifications, but not high coding gain. By dynamically switching between these parametric and non-parametric methods, we can obtain high coding gain and high quality compressed-domain modifications. In this system, we strive for a scalable range of low bitrates (20 to 40 kbps) while allowing for a large audio bandwidth (32 kHz sampling rate) and high quality compressed-domain time-scale modifications.

To achieve good data compression rates and high quality wideband modifications, we 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 [4], a signal which models all attack transients present using transform coding, and a Bark-band noise signal [5]. Each of these three signals can be individually quantized using psychoacoustic principles pertaining to each representation.

Transform coding is used for the transients because neither sinusoidal modeling nor noise modeling are able to accurately encode the attack transient 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 representations. 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 (slowed) time-scaled polyphonic music with percussion or drums, this results in slowed harmonic instruments and voice, with the drums still having *sharp* attacks [6].

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 [4] between 0 and $f_{\text{tonal}}(t)$ kHz. The time-varying ceiling in frequency, $f_{\text{tonal}}(t)$, dictates the highest possible frequency that will be represented by sinusoidal modeling. The sinusoidal modeling and quantization 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 [5], to be summarized in Section 5. Therefore, between 0 and $f_{\text{tonal}}(t)$ kHz, the non-transient signal model consists of sines and noise. Between $f_{\text{tonal}}(t)$ and $f_s/2$ kHz, there is only noise modeling. 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. In addition, a transient detector examines both rising energies in the original signal and in the sinusoidal residual signal to locate times to switch between parametric and transform coding [1, 6].

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 [3, 7]. In order to reduce pre-echo artifacts due to long analysis frame lengths, the input signal is initially split into four $2\times$ oversampled multiresolution octave-spaced signals [4]. The octave-band filter bank is oversampled to suppress any inter-octave aliasing below audibility. Parameter estimation is performed individually on each octave signal; the lowest octave has good frequency resolution, but poor time resolution. The highest octave has good time resolution, but has poor frequency resolution. The reasoning behind this system is that due to the near-logarithmic perception of pitch, frequency resolution is more important at lower frequencies than at higher frequencies.

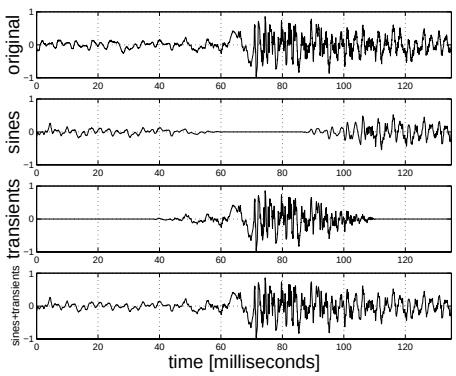

Figure 1: *In the top plot, the original signal is shown containing speech and a bass drum hit at time=65 milliseconds. The second plot shows the synthesized multiresolution sinusoids, which are faded out during the transient. The third plot shows the transform-coded transient, which is the residual 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.*

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 locally stationary 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 allowed 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. However, 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 1, 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 alignment, explicit phase information is transmitted for the sinusoids at least one frame before and after the transient region. Cubic-polynomial phase interpolation [7] is used only during this region to assure correct phase alignment. For more details, see [6].

3.2. Sinusoidal Parameter Quantization

With all the multiresolution sinusoidal parameters estimated, the next issue is to decide which estimated sinusoids to keep and which to eliminate. Ideally, any estimated sinusoids attempting to model noise processes are eliminated; the energy of these eliminated, falsely estimated sinusoids will be later represented more efficiently by the Bark-band noise modeling algorithm of Section 5. The remaining sinusoids are efficiently quantized, as will be shown soon. Exactly how many sinusoids to keep, and then how coarsely or finely to quantize them is also a trade-off between quality and bitrate. For those who desire modification quality more than data compression, e.g., for musical purposes, somewhat more sinusoids with finer quantization will be used. For those who desire data reduction more, methods will be described for gracefully lowering the data rate while still maintaining reasonable quality.

The set of steps for sinusoidal selection and quantization are graphically shown in Figure 2, and will now be described in more detail:

3.2.1. Psychoacoustic Model

The psychoacoustic model examines the perceptual relevance of each of the estimated multiresolution sinusoids. First, the masking threshold of the original input audio signal is computed using the methods described in the MPEG-2 Audio specification [8]. Theoretically, any audio source whose magnitude at a given frequency is of less magnitude than the masking threshold at that frequency will be inaudible. The ratio of the magnitude of a sinusoid to the masking threshold at its frequency is called the Signal-to-Masking ratio (SMR). In addition to the SMR for each sinusoid, the psychoacoustic model also computes a tonality ceiling, f_{tonal}, for each analysis frame. Roughly, it attempts to approximate the maximum frequency at which sinusoids should be retained. Any sinusoids above this time-varying frequency ceiling should be simply discarded, and later be modeled as only noise. More on this will be discussed in Section 3.2.4.

3.2.2. Eliminate Completely Masked Sinusoids

Some individual estimated sinusoids may have been erroneously detected due to sidelobe detection or errors in the algorithm. To quickly eliminate these erroneous sinusoids, any individual sinusoid with a greatly negative SMR is eliminated from the representation. Those sinusoids with a negative but near zero SMR will be kept at least until step (D) in Figure 2, as described in Section 3.2.5, since it might be associated with a quiet but stable sinusoidal trajectory.

3.2.3. Sinusoidal Tracking

If certain sinusoids are present at close to the same frequency and amplitude over a series of analysis frames, these sinusoids are *tracked*, and placed into a single *trajectory* [7]. Longer trajectories are usually associated with more *stable* harmonics. Because the interframe amplitude and frequency deviations are inherently limited by the tracking algorithm, the sinusoidal information can be efficiently encoded. The length of the trajectory will be used later for both sinusoidal selection and quantization.

3.2.4. Tonality Limits

In this step, all sinusoidal trajectories whose maximum frequency is above a certain $f_{\text{tonal}}(t)$ will be eliminated. The f_{tonal} parameter changes every analysis frame at the lowest octave, which is approximately every 23 msec. The basic concept of this step is that the energy of all sinusoidal trajectories whose frequencies are higher than $f_{\text{tonal}}(t)$ can be effectively modeled as noise with negligibly audible artifacts. All energy below this time-varying limit will be modeled as sines plus a residual noise model. This is in contrast to the fixed limit of $f_{\text{tonal}} = 5kHz$ presented in

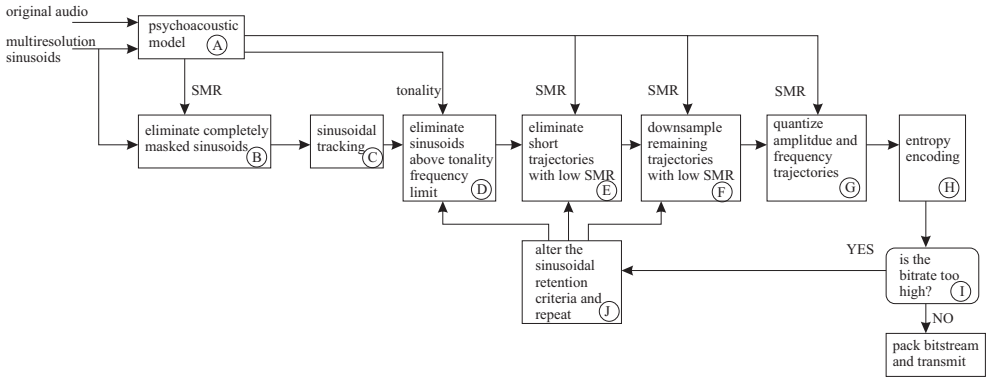

Figure 2: *Iteration loops of sinusoidal parameter selection and quantization process.*

earlier work [1]. While a fixed limit of 5 kHz worked well for most popular polyphonic music, it fared poorly for monophonic and strongly harmonic signals whose high frequency harmonics were insufficiently modeled by Bark-band noise only.

Other works have also successfully used noise modeling in speech and audio coding. In early speech coding techniques called Multiband Excitation Coding [9], certain spectral regions were either modeled by sines or by noise. In a separate speech modification method, all frequencies above a certain fixed limit are encoded as LPC-modeled noise [10]. Noise modeling is exclusively used in certain *noisy* frequency regions in place of transform coding in a audio coding method called Perceptual Noise Substitution in the MPEG-4 Audio specification [11].

3.2.5. Eliminate Noisy Sinusoidal Trajectories

The next goal is to eliminate all sinusoidal trajectories attempting to model noise processes. To perform this, two metrics are used: sinusoidal trajectory length and trajectory time-averaged SMR [1]. The theory is that noisy sinusoidal trajectories are both short-lived and have low SMR. But, one does not want to eliminate short-lived true sinusoids; nor would one want to eliminate a stable, yet quiet harmonic of an instrument. By combining these two metrics, as seen in Figure 3, the only short-lived trajectories that are not eliminated are those with high SMR. Also, longer and more stable sinusoids will not be eliminated, even if their time-averaged SMR is considerably lower.

3.2.6. Temporal Subsampling of Trajectories

At this point, most of the noisy and perceptually irrelevant sinusoids have been eliminated from the representation. These next three stages involve lowering the data required to represent the remaining sinusoidal parameters in a perceptually meaningful manner. In this module, the data rate is lowered by reducing the temporal resolution of the sinusoidal trajectories. After informal listening tests, it was found that reducing the temporal resolution by a factor of two for the sinusoidal trajectories with relatively low time-averaged SMR produced hardly any perceptible artifacts. In order to reduce the temporal resolution, only the odd or even time-indexed sinusoidal {amplitude,frequency} parameters are transmitted. At the decoder, each missing sinusoidal parameter in the subsampled trajectories is interpolated from its neighbors.

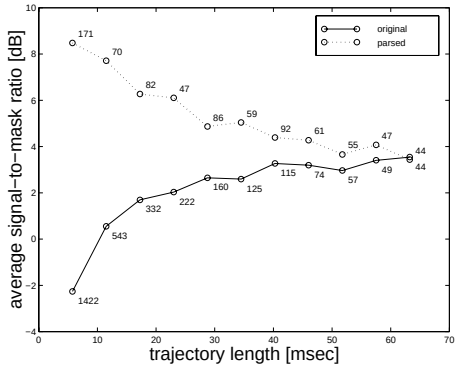

Figure 3: *SMR statistics of trajectory length of the original parameter estimation (lower solid curve) and the results of the trajectory selection process (upper dotted curve) as shown in Figure 2, module (E). The numbers next to each circled point on the curves show the total number of trajectories at the given length in the analyzed audio signal.*

3.2.7. Trajectory Parameter Quantization

In this module, each of the sinusoidal {amplitude,frequency} parameters are quantized to a near just noticeable difference (JND) scale. Quantizing each sinusoid to exactly a JND scale would have required far too many bits. Amplitude parameters are uniformly scalar quantized on a log-axis, with 1.5 dB of resolution. The frequencies are uniformly scalar quantized to a scale that closely follows the Bark scale. For more details, see [6]. After informal listening tests of natural audio input sources, no perceptual artifacts could be heard due to this stage of parameter quantization. In the future, the amplitude and frequencies can be quantized with a resolution that is a function of their individual SMR levels [12].

3.2.8. Entropy Coding

The basic unit of quantization is not the individual sinusoidal parameter of {amplitude,frequency}, but rather the sinusoidal trajectory as a whole. A trajectory can have as few as one analysis frame's worth of sinusoidal parameters, or as many as R frame's worth. The higher R becomes, the better coding gain one can achieve due to the correlation among sinusoidal parameters in the

same trajectory, as was discussed in Section 3.2.3. But, this comes at the expense of system latency, and a greater perceptual loss in case of data lost in transmission. In this system, simply the interframe differences of the amplitude and frequency parameters are computed, and then entropy encoded. In the future, more elaborate predictors could be used to further reduce the entropy of the trajectories' interframe parameters. Similarly, the initial amplitudes and frequencies of each trajectory are entropy encoded using separate Huffman tables.

3.2.9. Iterative Quantization Loops

Due to the various applications of such an audio representation, ranging from computer music high quality analysis / modifications / synthesis to low bitrate audio data compression, it is necessary to adapt the overall bit rate of the system. In this section, we will discuss only the methods for scaling the bitrate of the sinusoidal data, which is usually the majority of the bits utilized overall. In block (I) of Figure 2, the number of bits used in the entropy coding section, along with the header and side information, is counted. If the bitrate is too high, then block (J) lowers the f_{tonal} threshold of block (D), and raises the SMR thresholds for blocks (E),(F), and (G). Once these thresholds are altered, more sinusoids are eliminated from the representation, and the remaining ones are quantized more coarsely than before. For most audio inputs, the sinusoids required 8 to 20 kbps for high quality synthesis, depending on the tonality, signal complexity, and overall desired synthesis quality.

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 [6]. In order to reduce the overall bitrate, the time-width of the transient varies across frequencies. At low frequency, the width of the transform coded transient is 70 msec long, while at higher frequencies, the encoded transient width can get as short 20 msec long.

5. NOISE MODELING

Based on the work of [5], Bark-band noise modeling is used during non-transient segments of the audio signal. ¿From 0 to $f_{\text{tonal}}(t)$ kHz, the residual between the original and the synthesized multiresolution sinusoids is modeled as noise. From $f_{\text{tonal}}(t)$ to $f_s/2$ kHz, the original non-transient signal segments are modeled as Bark-band noise in order to reduce bitrate, as necessary, as was mentioned previously in Section 3.2.9. For most signals, $f_{\text{tonal}}(t)$ is usually in the range of 5 to 9 kHz. Bark-band noise modeling works by first performing a short-time windowed FFT upon the input signal. Then, FFT bins are grouped together in sets uniformly spaced on a Bark scale, then quantized, and transmitted. At the decoder, FFT bins are randomly generated using the quantized Bark-band gains. To further reduce the noise bitrates, the Bark-band gain envelopes themselves are quantized using line-segment approximation techniques [1].

6. CONCLUSION

In this paper, we described a low-bitrate audio compression system 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. Various methods have been discussed for automatically choosing which sinusoids should be retained and quantized, and which should be eliminated so that they are later modeled as noise.

7. REFERENCES

[1] S. Levine and J.O. Smith, "A switched parametric & transform audio coder," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Phoenix*, 1999.

[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," *J. Audio Eng. Soc.*, vol. 45, no. 10, October 1997.

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

[4] 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.

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

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

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

[8] ISE/IEC JTC 1/SC 29/WG 11, "ISO/IEC 11172-3: Information technology - coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s - part 3: Audio," 1993.

[9] D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 36, pp. 1223–1235, 1988.

[10] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Minneapolis*, pp. 550–553, 1993.

[11] J. Herre and D. Schulz, "Extending the MPEG-4 AAC codec by perceptual noise substitution," *Proc. of the 104th Convention of the Audio Engineering Society*, 1998, Preprint 4720.

[12] T. Verma and T. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Phoenix*, 1999.