Listener Evaluation of Reduction Strategies for Sinusoidal and Resonance Models

Amar Chaudhary[1,2] David Wessel[1] Lawrence A. Rowe[2]

[1]Center For New Music and Audio Technologies (CNMAT)
[2]Electrical Engineering and Computer Sciences
University of California, Berkeley, CA

This research explores strategies for dynamically reducing the size of the sound models used in additive synthesis and resonance modeling. Partials are ranked using measures of perceptual salience, and partials are removed in increasing order of salience to minimize the impact on human perception of the synthesized sound. Listening experiments were conducted to determine the effectiveness of strategies based on amplitude and masking. Using the results of these experiments, algorithms were developed based on the amplitude strategy.

Introduction

The widespread use of additive synthesis and resonance modeling in real-time music is frustrated by high computational requirements. Typical sinusoidal and resonance models contain hundreds of elements. Even though computers are now running above the 1GHz clock rate, it is still not possible to use many large models in polyphonic or multi-channel settings. This research explores strategies for dynamically reducing the size of sinusoidal and resonance models to reduce their computational load. It is part of a larger research effort to develop *perceptual scheduling* systems that automatically detect potential real-time performance failures in synthesis applications and automatically apply reductions to the synthesis models being used [1].

An acceptable strategy must be incremental, allowing perceptual quality to be gracefully traded for model size and synthesis execution time. Partials must be pruned from the models in a way that minimizes the impact on human perception of the synthesized sound. To this end, partials can be ranked by their *perceptual salience*, or impact on the quality of the sound. They can then be removed in increasing order of salience to reduce size and computation. Systems that prune partials according to perceptual salience when computational resources are scarce have been explored by Marks [2] and Haken [3]. Our research builds on previous work by exploring methods that not only dynamically prune the partials in real time, but also compute the measures of perceptual salience in real time. Such a system can adapt to timbres that are being modified in real time by a live performer.

A simple method for computing salience is the ranking and removal of partials by increasing amplitude (i.e., the softest partials have the least salience and the loudest partials the most). A measure of salience that better matches properties of the human auditory system is the relative *signal-to-mask ratio* (SMR) [4]. Although SMRs are a more accurate measure of perceptual salience than absolute amplitudes, computing SMRs and using them to prune partials is more computationally complex. For each pair of partials, the SMR is computed by taking the difference between the log amplitude of one partial to the value of "hat" function centered at the log amplitude and frequency of the second partial. If this value is negative, the first partial is masked by the second partial, and can be pruned without affecting the quality of the sound. Listening experiments were conducted to evaluate the performance of the two strategies and determine if the use of SMRs resulted in higher quality reductions to justify the added computational cost.

Measuring effects of reduction strategies

A group of 12 composers, performing musicians, and musical applications researchers were invited to participate in the listening experiments. Listeners were presented with pairs of sounds with one identified as the original and one identified as a possible reduction. Each listener was asked to rate any perceived difference on a scale from 1 to 5. The scoring system was chosen to correspond closely to both the 5-point Likert scale used in psychological surveys and the mean-opinion scores used in the audio/telephony community. Panelists were asked not to rate the absolute quality of the sounds (as in mean opinion scores), but rather the comparative quality between the reductions and the original. Early usability testing underscored the need to limit test pairs to comparisons where one is known to be the original.

Separate experiments were conducted for sinusoidal and resonance models. For the sinusoidal-modeling test, three models were chosen: a melody played on a suling flute, a rhythm played on a berimbao (a traditional Brazilian single-stringed instrument), and an except from a 1970 James Brown recording. The suling example contained 150 partials and lasted 7.9 seconds, the berimbao example contained 200 partials and lasted 3.5 seconds, and the James Brown example contained 250 partials and lasted 5.9 seconds. Reductions were made using both the amplitude and SMR strategies, with 8 reduced models between 3/4 and 1/64 of the original size for each strategy. Each listening session consisted of six randomly ordered trial groups representing each model and reduction strategy.

For the resonance-modeling test, three models with different sound qualities were chosen, a marimba, a plucked string bass, and a tam-tam, with 48 partials, 59 partials and 183 partials, respectively. Reductions were made using both strategies, 10 reduced models for each strategy between 3/4 and 1/32 of the original size.

Sinusoidal Model Results

All three models were rated consistently high (i.e., received ratings between 4 and 5), with a steep drop-off in perceptual quality for very reduced models (i.e., reduced below 1/4 the original size). In each case, the steep decline in quality can be attributed in part to the increase in the number of "birth and death events," or frames in a model in which a partial becomes present or absent. These events occur frequently when different partials are pruned from successive frames. These events cause loud artifacts when more salient partials are removed. The results for the suling model are illustrated in figure 1.

Although the results for each example exhibited similar trends, the onset and rate of the decline in quality does differ among the three examples. The suling model was rated highly when reduced to as few as 9 partials (i.e., 1/16 of the original size), and its ratings dropped steeply for higher reductions. The suling model is strongly harmonic with most of the energy concentrated in five partials, allowing very small reductions to retain most of the sound quality. More surprising is how well the "breathy" quality of the sound is maintained at higher reductions, implying that the noise energy is concentrated around a few frequencies. By contrast, the berimbao model has a richer spectrum and attack transients, so more of the sound is removed from the model at larger reductions. The James Brown example includes not only multiple notes but multiple instruments (i.e., voice, drums, electric guitar and bass, and horns) with complex overlapping spectra. While the voice and horns remained recognizable at higher reductions, the rhythm section instruments, and the percussion in particular, lost most of their spectral richness. There was also more change in the spectrum between successive frames because of the timbral variety and note changes. This caused the reduction algorithm to choose different partials in successive frames and produce more birth and death artifacts.

We can measure the "amount of sound removed" quantitatively as the ratio between the energies, or sum of amplitudes, in the reduced frame and the sum of amplitudes in the original frame:

$$\left(\sum_{i=1}^{N_R} A_i(n)\right) \Big/ \left(\sum_{i=1}^{N} A_i(n)\right) \qquad (1)$$

where N and N_R are the number of partials in the original and reduced frames, respectively, at discrete time n. From the results of the experiments, we conservatively estimate that such models can be reduced to approximately two-thirds of their original energy independent of the number of partials removed.

Another significant result is the strong similarity in perception of the amplitude and masking strategies. This result may seem at odds with the results of masking-based strategies used in perceptual encoders for compression. However, it is important to note that perceptual encoders remove masked frequencies from evenly-spaced samples in a full spectrum (i.e., the result of a Fourier Transform), while the frequencies in sinusoidal models are already distilled from full spectra via peak picking. On average only 25% of the partials in each model were masked

Resonance Model Results

The marimba resonance model exhibits a steep drop in quality below 7 partials (i.e., less than 1/8 of the original model size), which is similar to results for the sinusoidal models. However, the bass and tam-tam models show a more gradual decline in quality as partials are removed. Both the bass and tam-tam models received scores below 4 when more than half the partials were removed.

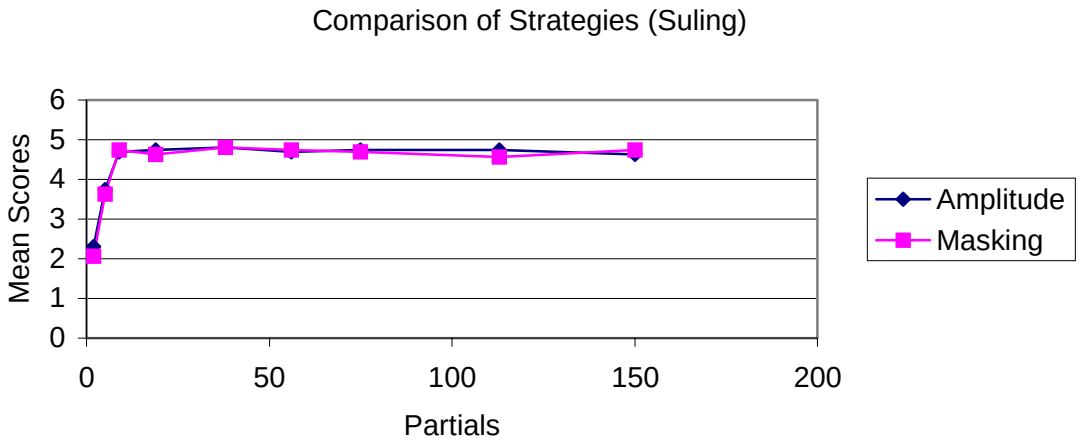

Figure 1. Results for reductions on the suling model. Both amplitude and masking strategies are displayed. In each graph, the *x* axis is the number of partials in each reduction and the *y* axis is the average user rating for each reduction

In the marimba model, only 7 partials contributed 93% of the initial energy (i.e., the sum of initial amplitudes). Most remaining partials represent the extremely brief attack at the beginning of the sound. Because no one partial has a strong spectral contribution to the attack, most can be removed without a strong effect on the sound. By contrast, the bass and tam-tam models are characterized by spectrally richer tones that decay over a longer period of time. In particular, most of the partials in the bass model contribute to a single harmonic series with additional partials as "beating pairs" for the harmonics below 900Hz. Removing the harmonic or beating frequencies greatly reduces the sound quality. In the tam-tam model, a large number of high-frequency partials contribute to the brightness and roughness of the sound. However, because each of these partials has a relatively low amplitude and each partial was likely to be masked by another nearby partial, they were more likely to be removed from the reductions. The removal of these partials gave the reductions below 50% a "duller" timbre.

The results of amplitude and masking strategies are again remarkably similar. While the reduced models of the marimba were nearly identical in both strategies, the equivalent-sized amplitude and masking of the bass and tam-tam only had 80% and 50% of partials in common. Although nearly 50% of the partials in the tam-tam were masked, most of these partials appear to have been masked by other low-amplitude partials, minimizing the difference in results between the two strategies.

Developing a Reduction Algorithm

Evaluating the reduction strategies requires weighing the perceived quality of reduced models against the computational cost that will be added to programs to perform the reductions. Amplitude-based reduction requires that all partials be sorted by descending amplitude, which is an $\theta(N\log N)$ operation where N is the number of partials. The masking algorithm described above requires $O(N^2)$ SMR calculations followed by $O(N\log N)$ operations to sort the unmasked partials. If few partials are masked, both operations approach their asymptotic costs. The fact that the amplitude and masking strategies showed nearly identical results leads us to choose the amplitude-based reduction strategy.

A real-time reduction algorithm must decide how many partials to remove without computationally expensive analysis or requiring that users perform their own listening experiments on their sound models. One method for estimating the reducibility of a model is the sum-of-amplitudes ratio described in equation 1. As stated above, each frame of a sinusoidal model can be reduced without significant degradation until the sum of amplitudes in the reduced frame is less than two thirds of the sum of amplitudes in the original frame. We therefore propose a reduction algorithm that removes the lowest-amplitude partials until the sum of the remaining amplitudes is only two thirds of the original.

In order to improve the quality of aggressive reductions of sinusoidal models, the issue of frequent births and deaths of partials must be addressed. Garcia and Pampin suggest discarding sinusoidal tracks shorter than a specified minimum if their average SMR measured over several frames is below a specified threshold [5]. In a real-time application with mutable sounds, we can only look at tracks backward in time. Moreover, we do not want to calculate averages over a large number of tracks because this approach increases computational cost and reduces temporal accuracy (e.g., the effect of a dramatic change in user input may be "smeared" over several frames in the reduced model because of averaging). We modify our algorithm by first adding to the amplitude of each partial in the current frame its amplitude in the previous frame if it was retained in the reduced version of the previous frame. This adds *N* steps to the algorithm.

Births and deaths are not an issue for resonance models, although we do want to modify the sinusoidal algorithm to include another temporal property, the decay rate. The results of the resonance model experiments emphasized the importance of partials with longer decay rates, particular in the bass and tam-tam models. Instead of using the initial amplitude of a resonance partial, we can compute its theoretical amplitude contribution over the duration of the sound as the ratio of initial amplitude to bandwidth:

$$\hat{A}_i = \int_t^\infty A_i e^{-\pi k_i t} = \frac{A_i}{\pi k_i} \qquad (2)$$

Substituting this formula for initial amplitude, we estimate that sum of contributions (i.e., $\hat{A}_i$ values) in a reduced model must be at least 90% of the original sum to maintain high quality. Although this sounds like a conservative estimate, it actually corresponds to a reduction by 7/8 for the marimba and by one half for the bass and tam-tam. The proposed reduction algorithm for resonance models prunes the partials with the smallest contributions until the sum of the remaining contributions is 90% of the original sum.

Applying Reductions to Multiple Simultaneous Models

In a larger sound-synthesis application, more than one model may be synthesized simultaneously. A computational reduction strategy must first select which models to reduce before then applying one of the reduction algorithms described above. The perceptual scheduling framework provides a means for multiple-model reductions by detecting potential real-time performance failures, estimating the computational bandwidth that would be saved applying reductions to each model and then selecting a subset of the models whose combined savings will avoid the real-time failure. This framework has been incorporated into the OpenSoundWorld real-time music and audio-processing environment [6]. The design and implementation of the framework, as well as its performance and evaluation in listening experiences, are described in a doctoral dissertation by Chaudhary [1].

Acknowledgements

This research was supported in large part by the National Science Foundation Graduate Fellowship Program and Gibson Music, Inc. We would also like to acknowledge advice and support of Adrian Freed, Matthew Wright, Richard Andrews, John Wawrzynek and Ervin Hafter.

References

1. Chaudhary, A., *Perceptual Scheduling in Real-time Music and Audio Applications*. Doctoral dissertation, Computer Science, University of California: Berkeley, CA., 2001.

2. Marks, M.A. Resource Allocation in an Additive Synthesis System for Audio Waveform Generation. *International Computer Music Conference*, San Francisco, CA., 1988.

3. Haken, L., Computational Methods for Real-time Fourier Synthesis. *IEEE Transactions on Signal Processing*, 1992. **40**(9): p. 2327-2329.

4. Terhardt, E., Psychoacoustic evaluation of musical sounds. *Perception and Psychophysics*, 1978. **23**: p. 483-492.

5. Garcia, G. and J. Pampin. Data Compression of Sinusoidal Modeling Parameters Based on Psychoacoustic Masking. *International Computer Music Conference*, Beijing, 1999.

6. Chaudhary, A., A. Freed, and M. Wright. An Open Architecture for Real-time Audio Processing Software. *107th AES Convention*, New York, 1999. `http://www.cnmat.berkeley.edu/OSW`.