FRACTAL SINUSOIDAL MODELLING FOR LOW BIT-RATE AUDIO CODING

Stuart K. Marks, Ruben Gonzalez

School of Information Technology
Griffith University
PMB 50, Gold Coast Mail Centre, QLD, 9726, Australia
s.marks@griffith.edu.au, r.gonzalez@griffith.edu.au

ABSTRACT

This paper proposes a fractal sinusoidal model that is able to reduce the bit-rate of sinusoidal model coders while achieving perceptually lossless quality. This is achieved by removing the redundancy between sinusoidal tracks through encoding similar tracks with the transformation between a template track and the original track. This paper proposes a transform that is able to capture the perceptual nature of sinusoidal tracks, and can be encoded efficiently. The results from our experiments show that the proposed fractal sinusoidal model coder is able to reduce the bit-rate of the sinusoidal model by roughly 30% while remaining perceptually lossless, while more aggressive modelling results in a reduction of around 60%, with minor quality degradation.

1 INTRODUCTION

Sinusoidal model audio coders [1] have been shown to be able to produce high-quality audio at low bit-rates [2,3,4]. These efforts were driven by the inability of other coding techniques, e.g. transform coders, to achieve good quality audio at these low bit-rates. In this paper we further reduce the bit-rate of sinusoidal model coders by applying fractal modelling to remove the self-similarities between sinusoidal tracks.

Fractal coding operates on objects; these objects are referred to as fractals and have self-similarity. The goal of fractal coding is to recreate objects by utilising the self-similarity to encode the data. With fractal image coding, a pattern is found that can be used to iteratively reconstruct the original object. This is called an Iterated Function System (IFS), consisting of an attractor (pattern) and a collage (a specification of the required iterations). The collage consists of iterations of affine transformations that are used to scale, rotate or stretch the [5].

Once a suitable attractor can be found fractal image codes enable flexible coding schemes that can produce low bit-rates. These characteristics are also beneficially for audio coding, however fractal audio coding has not been studied, except for the work of Wannamaker and Vrscay [6] that investigated the use of fractal coding to efficiently encode the wavelet coefficients for a wavelet audio coder.

This paper examines how fractal modelling can be applied to the encoding of mid-level audio representations to improve the performance of model coders. In this work we use sinusoidal tracks as our mid-level audio representation due to their high perceptual importance in model audio coding. It should be noted, however, that the same approach could be used with other mid-level representations, such as those used to represent transient and noise components of an audio signal.

Sinusoidal modelling generates sinusoidal tracks. Perceptually, tracks are objects that follow the evolution of a single partial or harmonic. Tracks are used with sinusoidal model as they improve the reconstruction quality while reducing the bit-rate [7]. Sinusoidal tracks are highly similar, and this similarity represents redundancy that can be removed to further reduce the bit-rate of sinusoidal model coders.

The approach taken in this paper is to perform fractal modelling on sinusoidal tracks to reduce the bitrate of audio coding. Fractal modelling encodes sinusoidal tracks as the transform from a template track. This approach has a high coding efficiency when the transform can be encoded with significantly fewer bits than the original track. This paper proposes such a transform.

The paper will begin with a description of the proposed fractal sinusoidal modelling technique, with particular emphasis on the transform that facilitates high coding efficiency. Then the results obtained by encoding audio samples using the fractal sinusoidal model coder will be presented, with the final section providing the conclusions of the paper.

2 FRACTAL SINUSOIDAL MODELLING

Fractal sinusoidal modelling reduces the bit-rate of audio coding by removing the redundancy that is present due to the similarity between sinusoidal tracks. This is achieved by encoding the transform from a template track to the original track.

We modelled the transform off sensible modifications of sinusoidal tracks, based on their perceptual nature. The transform provides mapping operators for frequency shift, amplitude gain, phase offset, time translation and time dilation of sinusoidal tracks. Figure 2.1 demonstrates how this can be achieved for two similar tracks in the frequency-time plane, with $track_b$ being a replica of $track_a$ that is delayed by Δt samples and is frequency shifted by Φf. Similar mappings exist in the amplitude-time and phase-time planes, as well as for track duration.

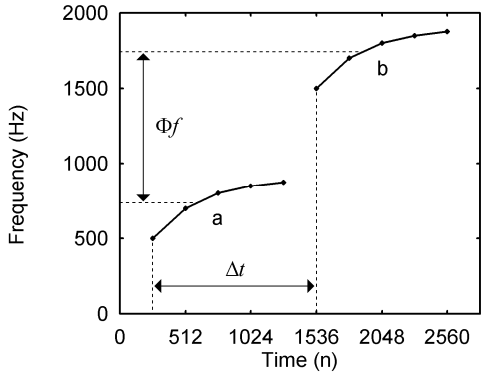

Figure 2.1. Fractal modelling of two similar sinusoidal tracks in the time-frequency plane.

These transform operators can be encoded cheaply, but do not provide perfect track reconstruction. As will be shown in section 3 this is not detrimental as sinusoidal tracks are not ideal representations themselves. They are based off a sequence of sinusoidal estimates, and a small amount of error will not be audible. It was also found that the error between the reconstructed track and the original could be determined by the similarity between the template and original tracks, providing a means for managing the modelling error.

The remainder of this section will examine the components of the fractal sinusoidal model in more detail. This includes the operators that define the transform and a similarity metric.

2.1 Transform

The five transform operators take scalar arguments which are calculated from the difference between the estimates from the template track, $track_a$, and the original track, $track_b$. The first operator is frequency shift, which maps the difference between tracks in the frequency plane. The argument is calculated using the average frequency estimate from each track as is shown in (1) where N_x is the number of estimates in track *x*, and $\hat{f}_{i,x}$ is the i^{th} frequency estimate in track *x*.

$$\Phi f_{a\to b} = \frac{\frac{1}{N_b}\sum_{i=0}^{N_b-1}\hat{f}_{i,b}}{\frac{1}{N_a}\sum_{i=0}^{N_a-1}\hat{f}_{i,a}} = \frac{\hat{f}_{mean,b}}{\hat{f}_{mean,a}} \quad (1)$$

The time translation operator allows tracks to occur at different times. The argument is calculated from the difference of the track onset times. This is demonstrated in (2) where $t_{0,x}$ is the time of the first estimate in track *x*.

$$\Delta t_{a\to b} = t_{0,b} - t_{0,a} = onset_b - onset_a \quad (2)$$

The duration of tracks can also vary; the time dilation operator maps this variation. The argument is determined by the ratio of track durations, as is shown in (3).

$$\Phi t_{a\to b} = \frac{\left(t_{N_b-1,b} - t_{0,b}\right)}{\left(t_{N_a-1,a} - t_{0,a}\right)} \quad (3)$$

Work conducted for the MPEG-4 high quality parametric coder has shown that the phase trajectory is characterised from the initial phase estimate and the sequence of frequency estimates [8]. Therefore the variation in phase trajectory between tracks can be determined from the initial phase offset. The phase offset operator does this precisely, with the argument being calculated using (4) where $\hat{\theta}_{0,x}$ is the first phase estimate of track x.

$$\Delta\theta_{a\to b} = \hat{\theta}_{0,b} - \hat{\theta}_{0,a} \quad (4)$$

The final operator accounts for the variation in amplitude between tracks. The argument for the amplitude gain operator uses the average amplitude estimate from each track to determine the argument; this is shown in (5).

$$\Phi A_{a\to b} = \frac{\frac{1}{N_b}\sum_{i=0}^{N_b-1}\hat{A}_{i,b}}{\frac{1}{N_a}\sum_{i=0}^{N_a-1}\hat{A}_{i,a}} = \frac{\hat{A}_{mean,b}}{\hat{A}_{mean,a}} \quad (5)$$

Using these operators a track can be recreated from a template track (6).

$$track_a \xrightarrow{T_{a\to b}(\Phi f,\Delta t,\Phi t,\Delta\theta,\Phi A)} \overline{track}_b \approx track_b \quad (6)$$

Tracks are recreated from the template track by copying each estimate and adjusting the parameters using the operator arguments. The amplitude estimates are scaled by ΦA, and the frequency estimates are scaled by Φf. The initial phase estimate is determined by adding the phase offset, $\Delta\theta$, to the initial phase estimate of the template track, then the sequence of phase estimates is predicted using the frequency estimates. The time of the estimates is shifted by Δt samples. This process continues until the track grows to the desired duration, as defined by Φt. This assumes that the template track is longer than the original track, so the time dilation must be in the range of $0 < \Phi t \le 1$.

The recreated track will be equivalent to the original if the two tracks are highly similar; otherwise the recreated track will not be equivalent to the original track. The next subsection presents a technique for determining the similarity between tracks that enables the track recreation error to be managed.

2.2 Similarity Metric

We measure the similarity between two tracks by using the perceptual measure proposed by Virtanen and Klapuri [9]. It measures the distance between normalised frequency and amplitude trajectories. This is beneficial as it automatically accounts for frequency shift and amplitude gain. The distance metric (7) measures the difference between frequency or amplitude estimates over the duration of the shortest track.

$$d_x(a,b) = \frac{1}{T}\sum_{t=0}^{T}\left(\frac{x_a(t)}{x_{mean,a}} - \frac{x_b(t)}{x_{mean,b}}\right)^2 \quad (7)$$

A similarity coefficient, σ, is then calculated from the distance of the frequency and amplitude trajectories, as is shown in (8). The similarity coefficient lies within the range of $0 \le \sigma \le 1$, with high coefficients indicating a high similarity between the tracks. The α, β and ρ coefficients are used to adjust the similarity measurement performance and bias. From experimentation it was found that an unbiased similarity coefficient, with $\alpha = 0.5$, performed best as information from the frequency and amplitude tracks are equally as important. It was also found that a setting $\beta = \rho = 10$ gave the best separation between similar and non-similar sinusoidal tracks.

$$\sigma(a,b) = \alpha e^{-\beta d_f(a,b)} + (1-\alpha)e^{-\rho d_a(a,b)} \quad (8)$$

Using this similarity metric the similarity between track combinations can be measured. Figure 2.2 provides the similarity coefficient measured for every track generated from a piano chord against the 5th track. It shows that tracks 5, 7 and 16 are similar, and thus can be modeled off each other. A similarity threshold, σ_T, can be used to determine when a track combination has adequate similarity, and will result in low-error track recreation.

For this paper a global search for similarities is employed, as our audio samples are relatively short at around 30 seconds each. Obviously a global search is not practical for longer audio samples; in this case a local search would be beneficial. From our experimentation, it appears that this would not be detrimental to performance as similar tracks are localised in time.

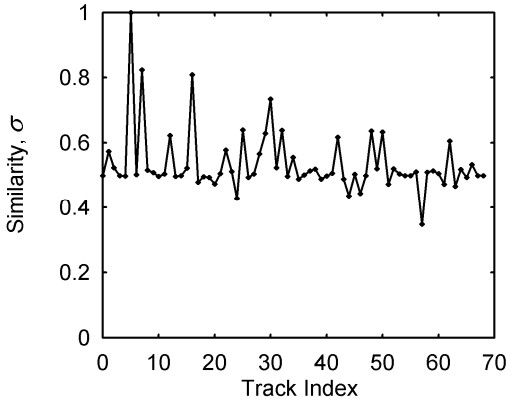

Figure 2.2. Similarity plot for track 5 of an audio sample of a piano chord. This demonstrates that tracks 5, 7 and 16 are similar.

3 RESULTS

To determine the performance of the fractal sinusoidal model coder a number of audio samples were encoded using the proposed coder. The coder uses multiresolution sinusoidal analysis [10] to generate the sinusoidal tracks. This includes the use of a CFB filter bank, sinusoidal estimation using quadratic interpolation, multiresolution sinusoidal tracking [11] and interpolating oscillators for synthesis.

The similarity metric is used to globally search all tracks for similarity. When the similarity is above a similarity threshold, σ_T, the track is encoded using the fractal model at a cost of 10 bytes (16 bits per operator argument). Otherwise the track is encoded using DPCM techniques as presented in [2,3,12].

The similarity threshold is the parameter that defines the rate-distortion performance of the fractal sinusoidal model coder, with high threshold values improving quality but providing little reduction in bit-rate, and low threshold values decreasing the quality

while significantly reducing the bit-rate. Our experiments investigated the performance of the coder against this parameter.

Figure 3.1 shows the number of tracks that are encoded using fractal modelling as the similarity threshold is adjusted. When more tracks are encoded using fractal modelling the reduction in bit-rate increases, as is illustrated in the right plot of Figure 3.1. Further reduction would be seen by entropy encoding the operator arguments, and remains as further work. The original bit-rate for the sinusoidal model coding is specified in Table 1 for each sample used.

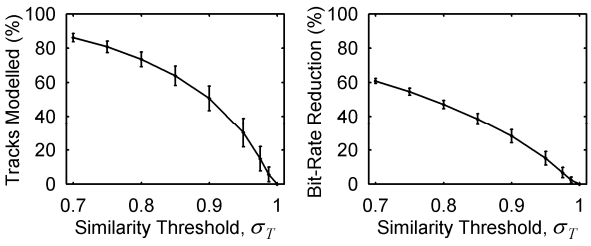

Figure 3.1. Percentage of tracks modelled (left) and bit-rate reduction (right) measured against the similarity threshold. Four stereo audio samples were used, and the error bars indicate the 95% confidence interval of the measurements.

The samples were reconstructed after fractal sinusoidal modelling to determine their perceptual quality. Perceptual quality experiments were conducted for each of the four samples listed in Table 1 using the ITU-R BS.1116-1 [13] test method. The reference signal was the reconstructed signal after sinusoidal modelling. The results for ten subjects are presented in Figure 3.2. The results indicate that the fractal sinusoidal model is able to provide lossless quality at $\sigma_T = 0.9$, with the subjects unable to differentiate between the original sinusoidal modelled and fractal modelled samples. The average reduction at this threshold was 28.19% across the four samples. More aggressive modelling, $\sigma_T \leq 0.8$, provides a slight reduction is quality, with the subjects being able to perceive the difference between the original and modelled samples. At these thresholds the bit-rate reduction reaches up to 60% on average.

Sample	Original	30%	60%
Jack Johnson	32.49	22.74	13.00
Jamiroquai	32.31	22.62	12.92
Led Zeppelin	77.68	54.38	31.07
Mozart	24.91	17.44	9.96

Table 1. The bitrate (Kbps per channel) for the original sinusoidal modelled samples, and the corresponding bit-rates for 30% and 60% bit-rate reduction. The high values for the Led Zeppelin sample are due to the large amount of transient signal energy present in this sample.

While it could be argued that a similar reduction in bit-rate could be achieved by limiting the number of encoded tracks, from our experiments this approach should be avoided as it creates audio which begins to sound synthetic due to the lost signal components. The benefit of the fractal sinusoidal model is that it provides a cheap method for encoding tracks, instead of removing tracks completely. The fractal sinusoidal model is able to provide a signal reconstruction that has perceptually lossless quality at low bit-rates.

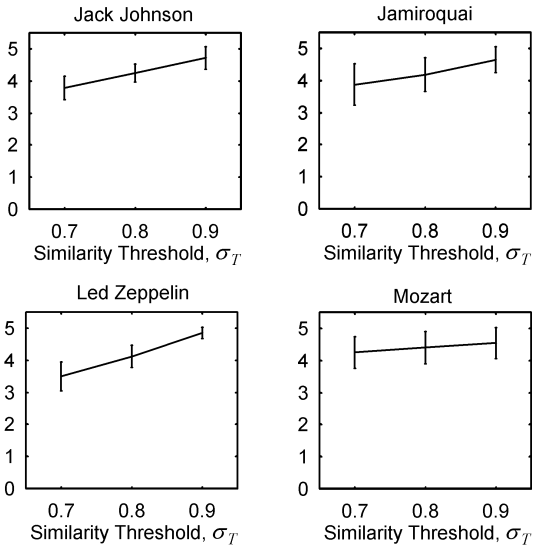

Figure 3.2. Mean perceptual quality of the reconstructed audio samples from the fractal sinusoidal model coder. The error bars represent the 95% confidence interval.

4 CONCLUSION

This paper presented a fractal sinusoidal model coder that is able to efficiently encode sinusoidal tracks by encoding the transformation from template tracks. A transform was presented that could be efficiently encoded, but is also capable of capturing the perceptual characteristics of the sinusoidal tracks. This resulted in roughly a 30% reduction in bit-rate while providing perceptually lossless quality for conservative modelling. While more aggressive modelling results in around 60% bit-rate reduction with minor quality degradation.

5 ACKNOWLEDGEMENTS

Australian Research Council's Spirt Scheme,
ActiveSky Inc.

6 REFERENCES

[1] McAulay, R. and Quatieri, T., "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 34, no. 4, pp. 744-754, Aug. 1986.

[2] Verma, T. and Meng, T., "A 6Kbps to 85Kbps scalable audio coder", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00*, vol. 2, pp. 877-880, 2000.

[3] Hamdy, K., Ali, A. and Tewfik, A., "Low bit rate high quality audio with combined harmonic and wavelet representations", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'96*, May 1996.

[4] Purnhagen, H. and Meine, N., "HILN - The MPEG-4 Parametric Audio Coding Tools", *IEEE International Symposium on Circuits and Systems, ISCAS 2000*, Geneva, May 2000.

[5] Wohlberg, B. and de Jagerm G., "A Review of the Fractal Image Coding Literature", *IEEE Transactions on image Processing*, vol. 8, no. 12, Dec. 1999.

[6] Wannamaker, R. and Vrscay, E., "Fractal wavelet compression of audio signals", *Journal of the Audio Engineering Society*, vol. 45, pp. 540-553, Jul. 1997.

[7] Verma, T., "A perceptually based audio signal model with application to scalable audio compression", *PhD Thesis*, Stanford University, Oct. 1999.

[8] Schuilers, E., et al., "Advances in parametric coding for high-quality audio", *Proceedings of IEEE Bebelux workshop on model based processing and coding of audio*, Leuven, Belgium, MPCA-2002, Nov. 2002.

[9] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modelling", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00*, vol. 2, pp. 765-768, Jun. 2000.

[10] Levine, S., Verma, T. and Smith, J., "Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio", *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 1997.

[11] Marks, S. and Gonzalez, R., "Techniques for improving the accuracy of sinusoidal tracking", Proceedings of IASTED European Conference on Internet Multimedia Systems and Applications IMSA EuroIMSA 2005, Feb. 2005.

[12] Marchand, S., "Compression of sinusoidal modeling parameters", Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy, Dec. 2000.

[13] "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", ITU Recommendation BS.1116-1, http://www.itu.org