2001

Audio coding using sorted sinusoidal parameters

Mohammed Raad
University of Wollongong, mraad@uow.edu.au

I. Burnett
University of Wollongong, ianb@uow.edu.au

Publication Details

This article was originally published as: Raad, M & Burnett, I, Audio coding using sorted sinusoidal parameters, The 2001 IEEE International Symposium on Circuits and Systems, (ISCAS 2001), 6-9 May 2001, 2, 401-401. Copyright IEEE 2001.

Audio coding using sorted sinusoidal parameters

Abstract
This paper describes a new audio coding scheme based on sinusoidal coding of signals. Sinusoidal coding permits the representation of a given signal through the summation of sinusoids. The parameters of the sinusoids (the amplitudes, phases and frequencies) are transmitted to allow the signal reconstruction. In the proposed scheme, the sinusoidal parameters are sorted according to energy content and perceptual significance. The most significant parameters are transmitted first allowing the use of only a small set of the parameters for signal reconstruction. The proposed scheme incurs a low delay and uses a 20 ms frame length. Results show that the coder operating at a mean rate of 39 kb/s, performs favorably in comparison with the MPEG-4 coder at 42 kb/s.

Keywords
audio coding, delays, signal reconstruction

Disciplines
Physical Sciences and Mathematics

Publication Details
This article was originally published as: Raad, M & Burnett, I, Audio coding using sorted sinusoidal parameters, The 2001 IEEE International Symposium on Circuits and Systems, (ISCAS 2001), 6-9 May 2001, 2, 401-401.

Audio coding using sorted sinusoidal parameters

M. Raad and I.S.Burnett
School of Electrical, Computer and Telecommunications Engineering.
University of Wollongong, Northfields Ave Wollongong NSW 2522, Australia.
mr10@uow.edu.au

ABSTRACT

This paper describes a new audio coding scheme based on sinusoidal coding of signals. Sinusoidal coding permits the representation of a given signal through the summation of sinusoids. The parameters of the sinusoids (the amplitudes, phases and frequencies) are transmitted to allow the signal reconstruction. In the proposed scheme, the sinusoidal parameters are sorted according to energy content and perceptual significance. The most significant parameters are transmitted first allowing the use of only a small set of the parameters for signal reconstruction. The proposed scheme incurs a low delay and uses a 20ms frame length. Results show that the coder operating at a mean rate of 39 kb/s, performs favorably in comparison with the MPEG-4 coder at 42 kb/s.

1. INTRODUCTION

Audio coding aims to reduce the bandwidth required for storage and transmission of a digitized audio signal. Nominally, single channel CD quality audio requires a transmission rate of 706 kb/s [5]. The reduction of this bit rate has been the impetus behind the introduction of standards such as MPEG-1, 2 and 4 as well as products such as the Dolby AC-2 and AC-3 systems for digital audio compression and transmission [5].

This paper focuses on reducing the bit rate required for the transmission of audio by using a different approach to the existing schemes. Sinusoidal coding of the audio signal is combined with a sorting technique to allow better quality audio to be produced at low bit rates. The sorting technique emphasizes the signal frequency components that have more effect on the perceptual quality of the audio signal and de-emphasises those components that have little or no effect on signal quality.

This paper borrows techniques used in both stereo audio coding as well as sinusoidal coding. However, it is primarily aimed at offering audio enhancements to wireless applications (such as GSM mobile telephony). Thus the sinusoidal coding technique utilises a short, fixed, frame length, setting it aside from typical sinusoidal coders that use a variable frame length [1] [2].

2. SINUSOIDAL AUDIO CODING

Sinusoidal coding of signals has been used in the past to develop both speech [1] and audio (music) coders [2]. The following sections describe the general principles behind sinusoidal coding and the modifications made to the general model for the work of this paper.

2.1 The general model

The principle behind sinusoidal coding is the Fourier representation of a signal whereby a given signal is represented by its sinusoidal components such that:

$$s[n]=\sum_{i=1}^{L} A_i \cos(\omega_i n+\phi_i) \quad n=0,\ldots,N-1 \qquad (1)$$

where A_i, ω_i and ϕ_i are the amplitudes, frequencies and phases of the sinusoidal components respectively and s is a selected frame, of length N, of the original audio signal. In the proceeding sections, the amplitudes, frequencies and phases will be regularly referred to as the parameters of the sinusoidal model.

There are two popular techniques for deriving the sinusoidal parameters: direct use of the Discrete Short Time Fourier Transform (DSTFT) [1] or the use of Analysis-By-Synthesis (ABS) [2]. Using the Fourier transform method, the sinusoidal parameters are determined from the DSTFT parameters such that:

$$A_i=\sqrt{(a_i^2+b_i^2)},\ \phi_i=\arctan\left(\frac{-b_i}{a_i}\right) \text{ and } \omega_i=\frac{2\pi i}{N} \qquad (2)$$

where a_i and b_i are the DSTFT coefficients and N is the total number of transform coefficients.

To allow the DSTFT of the signal to be performed, the signal is divided into frames by the use of a satisfactory window. Typically, window functions such as the Hamming or Hanning windows are used [1][2][3] and the frames are overlapped by up to half the given frame length. The length of the window is normally kept at 2.5 times the length of the pitch period of the signal. The pitch period is regularly estimated and updated, resulting in a variable length frame coder.

The reconstruction of the signal is performed by employing the overlap-add technique to avoid discontinuities at frame boundaries. That is, the synthesised signal is given by:

$$\hat{s}[n]=W_s[n]\hat{s}^{k-1}[n]+W_s[n-T]\hat{s}^{k}[n-T] \qquad (3)$$

where $\hat{s}^k$ is the synthesised k^{th} frame, W_s is the reconstruction window and T is the overlap of the consecutive frames (in samples) [1]. It has been shown that some performance gains may be obtained by ensuring that the analysis and synthesis windows are identical [3].

2.2 Modifications made

The general model described in section 2.1 was modified slightly for the work described in this paper. Firstly, a short, fixed frame is used which is more fitting for a coder to be used for continuous transmission purposes. In choosing the frame length of the coder, a lead was taken from the GSM speech coder which utilises a 20ms frame length [6]. The use of a 20ms frame length means that the audio coder proposed is a short frame length coder.

Frame selection is carried out by the use of overlapping windows. At a sampling frequency of 44.1 kHz (the sampling frequency of the CD) a 20ms frame corresponds to

0-7803-6685-9/01/$10.00©2001 IEEE

approximately 880 samples per frame. The window used in this case was obtained from [8] and is given by:

$$W[n] = \sin\left((n+0.5)\frac{\pi}{N}\right) \quad 0 \le n \le N-1 \qquad (4)$$

W is a perfect reconstruction window of length N and so is used for both analysis and synthesis. This window is also used because of its sharp main lobe and low side lobe frequency response. The narrow main lobe ensures minimal "smoothing" effects in the frequency domain and the low side lobes reduce the potential of "frequency leakage". The use of a perfect reconstruction window is consistent with such coders as AC-3 and MPEG-4 [5].

3. PROPOSED ARCHITECTURE

Figure 1 illustrates the proposed audio coder and decoder.

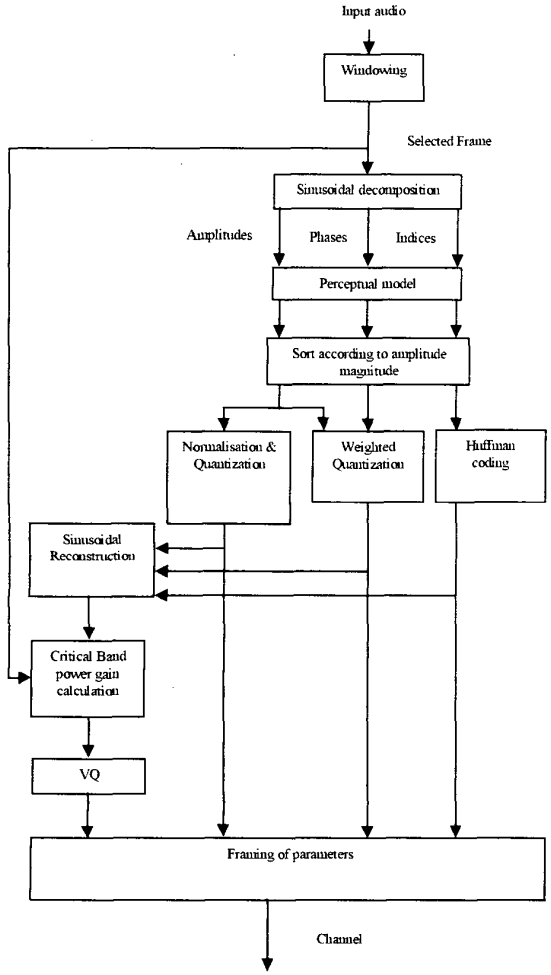

Figure 1 (a) Proposed encoder architecture.

The proposed architecture is built around the sorting of the amplitudes according to energy content. We thus consider the sorting technique first in this discussion.

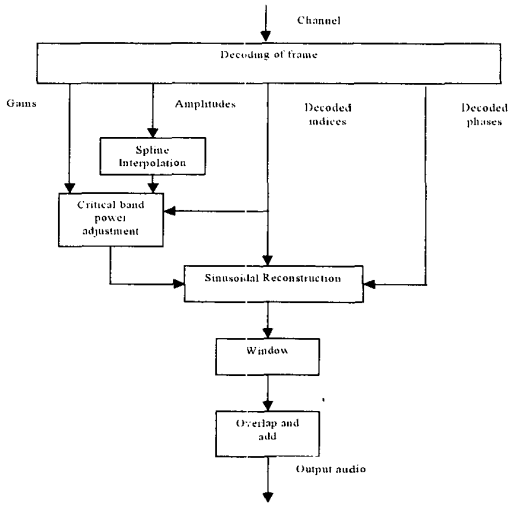

Figure 1(b) The proposed decoder

3.1 Sorting the parameters

It is clear from equation (1) that the energy content of each of the samples in a frame is determined by the sum of the square of the amplitude magnitudes. As this is the case, the amplitudes are sorted in order of decreasing magnitude. The sorting of these amplitudes allows the coder to concentrate on the parameters that contribute most to the reconstruction of the original signal. The sorting also provides the added advantage of producing a monotonic relationship between the ordered amplitude magnitudes. This relationship permits the modeling of the amplitudes by the use of either a monotonic decreasing function or by the use of interpolation. A further advantage is that, having arranged the amplitudes in a manner that determines the amplitudes' relative importance, scalable audio coding with a smooth increase (or decrease) in quality is a possibility.

Using the sorting scheme described, it was found, through informal listening tests, that only fifty sets of parameters (i.e. amplitudes, phases and indices) need be used (out of a complete 441 in the frame selected) to achieve good quality synthesised audio.

Simply sorting the amplitudes in terms of energy content does not take advantage of the perceptual redundancies in the signal. To improve transmission rates while maintaining quality a perceptual model is employed to remove the perceptually insignificant components from the model before the sorting of the amplitudes. We now explore the use of simultaneous masking as a perceptual model for this selection process.

3.2 The perceptual model

Sorting the sinusoids according to energy content to determine the relative importance of each sinusoid is an effective technique when the signal to be coded is highly tonal. However, this technique will not perform as well for non-tonal signals as the

energy content of non-tonal signals is distributed over more frequencies than tonal signals.

It is known from auditory theory that when numerous sounds reach the ear simultaneously, a number of them may be masked [5]. In this work, the masking effect is used to determine which frequency components are masked (according to the original signal's frequency representation). In particular, we seek to eliminate a number of high-energy components in the same critical frequency band. This allows either a reduction in the number of components that need transmission or an increase in sound quality using the same number of components. The technique used to calculate the simultaneous masking curve for each frame is the same as that presented by Johnston [7].

The application of the masking model differs from the technique used in, for example, MPEG-4 and AC-3 where the model determines a limit for the quantization noise [5]. Here, the perceptual model is being used in the selection of the frequency components to be transmitted first.

Using the simultaneous masking technique, it was found, through informal listening tests, that thirty-five sinusoids can be used to produce comparable quality audio to fifty sinusoids selected purely on energy content.

3.3 Frequency domain gains

As mentioned in the preceding two subsections, only a small number of sinusoids need be used from the complete set of sinusoids to produce the synthesised audio. A consequence of using so few sinusoids from the original set is the reduction of overall energy content in the signal. To counter significant energy loss, twenty five frequency domain gains are used, each corresponding to a critical frequency band as given in [5]. These gains adjust the average energy of the synthesised signal in each critical band to ensure the energy in each band is approximately the same for the synthesised and original signal.

In practice, it was found that more weight should be applied to the low frequency gains than the high frequency gains. Without this weighting high frequency "scratches" may occur in the synthesised audio when the gains are used quantized. These scratches are undesirable and the weighting is carried out by applying a power law. That is, the actual gains used are related to the calculated gains by:

$$\Gamma_{f_\ell} = (\Gamma_{i_\ell})^{x_\ell} \quad 0 \le \ell \le 24 \qquad (5)$$

Γ_{f_ℓ} and Γ_{i_ℓ} are the final and initial gain vectors respectively and x is a weighting vector varying between 0 and 1. The power law technique allows the gains to be adjusted between their original values and 1. This permits for the adjustment of sinusoids belonging to the low frequency critical bands whilst leaving others unchanged. The initial gains are calculated by taking the ratio of the total energy in a given critical band before modeling the signal with a reduced number of sinusoids to the final total energy in that particular band after the modeling. The critical band gains have been found to contribute significantly to the perceived quality of the synthesised signal.

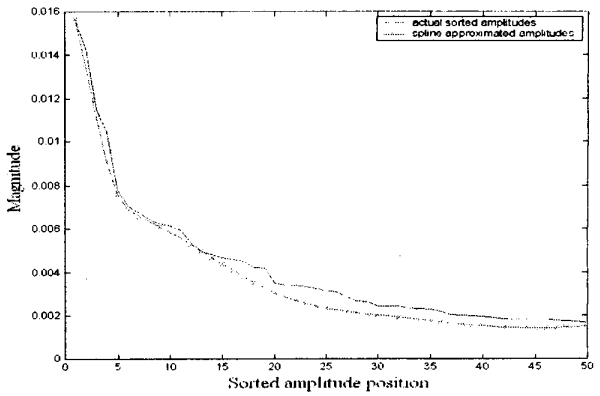

Figure 2 Sorted amplitudes and their spline interpolation model.

4. QUANTIZATION

The sorting of the sinusoids according to energy content allows preferential transmission of the parameters as well as the exploitation of the relationship between consecutive amplitudes.

4.1 Quantization of the amplitudes

The sorted amplitudes have a monotonic relationship, due to the sorting (see Figure 2). The gradient of the amplitude magnitudes is related to the signal being coded, with tonal signals having a much higher rate of decrease than non-tonal signals. The top curve in Figure 2 is that of the sorted amplitudes for a non-tonal signal.

Two approaches for quantizing the amplitudes were considered. The first models the amplitudes using a smooth function, such as an exponential, and the second uses spline interpolation. In the case of exponential modeling, the top curve in Figure 2 is modeled by attempting to fit an exponential function to it. On the other hand, the interpolation technique involves the selection of a number of amplitudes, quantizing those amplitudes and using a spline interpolator between the quantized amplitudes. Figure 2 shows how closely the spline interpolator approximates the actual amplitudes (the bottom curve) using 5 bits per amplitude. Note that the latter technique relies on the accuracy of the quantization and the position of the selected amplitudes. Figure 3 shows MSE results obtained for both techniques. The results in Figure 3 are averages and are displayed as percentages of the maximum amplitude magnitude for ease of comparison.

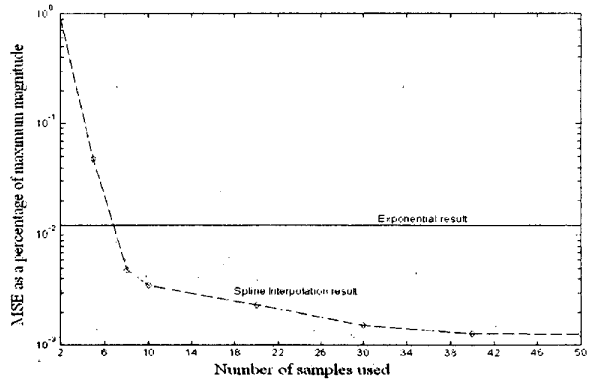

Figure 3 Average MSE results for selected number of amplitudes

In Figure 3 the effect of the quantization noise on the MSE results is evidenced by the non-zero MSE for 50 quantized amplitudes used to represent 50 actual amplitudes. Figure 3 shows that the use of 10 qauntized amplitudes performs well.

4.2 Quantizing the phases

In sinusoidal speech coders, the phase tends to be modeled rather than quantized, as in [1]. A different approach taken by [2] was to maintain the time domain envelope of the signal which implicitly contains phase information. It was found that, in this case, the phase is quite important, a result confirmed by [1]. In our coder, the phase is quantized using weighted scalar quantization, where the amplitudes are used as relative weights. That is, phases that are associated with the large amplitudes are quantized more accurately than phases associated with the smaller amplitudes. The quantizer currently uses four bits per phase for the phases associated with the ten largest amplitudes, three bits for those associated with the next fourteen amplitudes and two bits per phase for the remainder. This distribution was developed after experimentation with a wide variety of audio signals.

4.3 Quantizing the indices

The indices indicate the spectral position of the sinusoids before sorting. As the sinusoids have been re-arranged, the indices are required by the decoder for correct synthesis. To avoid distortion, the indices must be coded losslessly. A total of 441 sinusoids are used; thus nine bits per index would be required for fixed length loss-less coding. However, the first order entropy of the indices was found to be 4.24, as some frequencies have a higher probability of appearing as part of the most energetic fifty sinusoids than others. Thus the theoretical limit for coding the indices is 4.24 bits per index [5]. Using this result, a Huffman code was designed for the indices based on the probability of appearance in the most energetic fifty sinusoids. The Huffman code resulted in a variable length code set for the indices with a mean length of 6.089 bits per index, a reduction of 32%.

4.4 Quantizing the gains

The vast majority of the critical band energy matching gains were found to be distributed between one and two. A ten bit codebook was trained to quantize the gains with each gain vector having 25 elements. The use of this codebook in combination with the weighting of the gains produced insignificant distortion on the synthesised audio.

5. RESULTS

Using a total mean bit rate of 38.9 kb/s (i.e. 35 sinusoids), the proposed coder was found to synthesise audio with comparable quality to the MPEG-4 transform coder at 42kb/s. It was notable that the proposed coder performed better at these rates than the MPEG coder for wide band speech test files. The subjects of the tests are the SQAM files obtained from [4]. Table 1 shows some Segmental Signal to Noise Ratio (SegSNR) results for the proposed coder at 38.9 kb/s. The SegSNR results obtained indicate that the proposed coder performs reasonably well at medium range bit rates (it should be noted that a SegSNR of 10 dB and above was found to indicate good perceptual quality). As expected, the tonal signals tend to perform best (as can be seen from the Glockenspiel and Horn results), whereas the most non-tonal signal of the set (Harpsichord) performs the worst (section 3.1). The SegSNR results were found to be indicative of perceived quality in informal listening tests.

SIGNAL	DESCRIPTION	SEGSNR
S1	Bass	13.417
S3	Glockenspiel	15.305
S5	Harpsichord	7.280
S7	Qaurtet	14.094
S9	Eng F speech	10.863
S12	Eng M speech	10.683
S6	Horn	17.293

Table 1 Some SegSNR results for the proposed coder using 50 sinusoids

6. CONCLUSION

A sinusoidal coder employing a different technique for transmitting the amplitudes has been presented. The sinusoids are rearranged according to each sinusoid's energy contribution to the reconstruction of the original signal. A perceptual model is also used to enhance the selection process and ensure the selection of perceptually significant information. The coder presented uses a short frame length and minimal algorithmic delay in the coding of the audio signal. The audio coder presented has been found to perform well at bit rates close to the 40 kb/s mark.

7. ACKNOWLEDGEMENTS

M.Raad is in receipt of an Australian Postgraduate Award (Industry) and a Motorola (Australia) Partnerships in research Grant. Whisper Laboratories is funded by Motorola and the Australian Research Council.

8. REFERENCES

[1] McAulay, R.J. Quatieri, T.F. "Sinusoidal coding" Chapter4 in "Speech coding and synthesis" edited by Kleijn, W.B. and Paliwal, K.K. Netherlands: Elsevier publishing, 1995.

[2] George, E.B. and Smith, J.T. "Analysis-by-Synthesis/Overlap-add sinusoidal coding applied to the analysis and synthesis of musical tones" J. Audio Eng. Soc., Vol. 40, No. 6, pp. 497-516. June 1992.

[3] Vos, K. Vafin, R. Heusdens, R and Kleijn, W.B. "High-quality consistent analysis-synthesis in sinusoidal coding", 17th AES international conference on High Quality Audio coding, pre-print version.

[4]http://www.tnt.uni-hannover.de/project/mpeg/audio/#mpeg4

[5] Gibson, J.D. Berger, T. Lookabaugh, T. Lindbergh, D. Baker, R.L. "Digital compression for multimedia". USA: Morgan Kaufmann Publishers, Inc. 1998.

[6] Eberspacker, J. Vogel, H-J. "GSM switching, services and protocols" Great Britain: John Wiley & Sons ltd. 1999.

[7] Johnston, J.D. "Transform coding of audio signals using perceptual noise criteria" IEEE Journal on selected areas in communications vol. 6, No. 2, pp. 314-323 February 1988.

[8] Malvar, H.S. "Signal Processing with Lapped Transforms". USA: Artech House, 1992.